平成26年度 卒業論文

変数を見直したベイジアンネットワークによる
要注意学生の発見手法に関する研究

指導教員
舟橋　健司　准教授
伊藤　宏隆　助教

名古屋工業大学　工学部　情報工学科
平成23年度入学　23115014番

名前　稲垣　諒

目 次

第1章　はじめに

名古屋工業大学では, 双方向型教育支援システムの構築を目的として, IC カード出欠管理システムと Course Management System（コースマネジメントシステム：以下 CMS）を導入している [1]. IC カード出欠管理システムは,IC カード化された学生証を, 入室時と退出時に教室に設置されている IC カードリーダーにかざすことで, 授業の出席をとることができる. この情報は教員が Web 上で参照することができ, 学生の最終評価の指標などにも活用されている. CMS は, 情報技術やインターネットを使った e-Leaning を支援するシステムである. 教材の作成支援や資料の配布, 課題の提出管理, 小テストの実施, 受講者の管理を Web 上で行うことができる. IC カード出欠管理システムや CMS は学生の情報を電子データとして蓄積する. 電子データとすることで, 大量のデータの保持や参照スピードの向上に大きく寄与した. 近年ではそれだけではなく, データマイニングによって新たな知識や傾向を見つけようとしている. データマイニングとは, 大量のデータの中から有用な知識を見つける技術であり, マーケティングや株価予測などの商業や, 臨床データに基づいた病気の経過や薬の効果の予測の医療などの分野では実用的に用いられている.

教育現場におけるデータマイニングの活用方法として, 学生に関するデータから一人ひとりの修学傾向を読み取り, 何かしらの学習指導を行うという提案がされている. 過去の関連研究では, 講義の出席状況や課題提出状況から学生の成績を予測したもの [2] や, 打刻データと成績から学生の学習レベルの予測をしたもの [3], 学生に対して行われる授業アンケートをもとに, 成績や授業評価の関係性を調査したもの [4] が挙げられる.

近年社会の多様化により大学生の性質も大きく変わってきている. それにともない大学では, 消極的な理由による退学者が目立ってきている. 大学生が退学する理由は, 家庭の経済的貧困や学生自身の病気や怪我などのどうしようもない場合や, 転学などの積極的理由による場合があげられる. また中には大学生活に馴染めない学生や, 真面目とは言えない学生が学校に来なくなってしまうことも多く指摘されている. さらに, 就職や大学院入試に失敗した学生が計画的な留年をする場合も少なからず存在する. なぜこのように退学してしまうのか, 原因究明のためにデータマイニングへの期待が高まっている.

現在, 先述した退学してしまう学生を助け出すため, 学生と教師が直接向かい合って, 学習面や生活面でのアドバイスや相談を行う指導方法が多くの大学でとられている [5]. しかしこの方法では, 一人の教員が多くの学生を指導する場合, 教師の負担が大きくなってしまい結果的に十分な指導が行えない可能性がある. また, 指導をするにも, 判断するデータがなければ指導そのものを行うことができない.

そこで, 過去の研究にて, 留年や退学をする学生を調査・分析し,「要注意学生」を定義し, この「要注意学生」を予測する研究が行われた [6]. 予測により学習指導者を絞ることで, 学習指導の時間的コストを削減している. さらに, 予測された学生は分析・調査によって定義

された「要注意学生」であるので, 指導の仕方も判断しやすい. この研究では, 成績の指標に Grade Point Average（以下 GPA）用いている. GPA が 1 年前期または 1 年後期で 1.0 を下回る学生は指導が必要であることは明白であるため, 必然的に学習指導対象とする. その他に 1 年前期または 1 年後期の GPA が 1.0 を上回りつつ, 今後留年または退学してしまう学生を「要注意学生」と定義して, 予測を行っている. また, この予測にはベイジアンネットワークを用いている. この手法により 122 人の学習指導対象者を挙げており,1 年次以降に留年または退学する学生の 81.4 %を, 全学生 338 人を指導する場合の約 3 分の 1 の時間的コストで発見できることを示しており, ベイジアンネットワークによる「要注意学生」の発見の有用性を示している.

しかしながら, 実際に指導やアドバイスを必要としているのは, 学校になじめない学生や学業に不安がある学生である. 「要注意学生」を一律に 1 年次前期・後期の GPA が 1.0 より大きく今後留年または退学する学生としてしまうと, 実際に指導やアドバイスを必要としている学生と, 転学や計画的な留年をする学生のような指導やアドバイスをあまり必要としていない学生が混合してしまう. これでは実際に指導やアドバイスが必要な学生が発見されずそのまま大学を去ってしまうかもしれない. そこで, 本研究では「一年次の GPA が 1.0 より高くかつ今後消極的理由により留年や退学する学生」と「要注意学生」を定義して発見・予測を行う. ここで使っている消極的理由とは, 先にも説明した「学校になじめない」や「学業に不安がある」などのような理由である. このように「要注意学生」を定義することで, 以前までより「要注意学生」となる学生の傾向をつかむことができるので, 発見・予測の精度が向上することが期待される.

本研究では成績データ, 打刻データ, 出欠データからベイジアンネットワークを用いて「要注意学生」の発見・予測を行う. 過去のある 2 つの年度 338 人を対象にしている. 成績データとは, 文字通り学生の授業の成績である. 成績から指標として成績別獲得数と GPA を採用した. 成績別獲得数とは, 秀や可といった成績をいくつ獲得したかの数である. GPA は総合的な GPA だけでなく, 理科や数学, 専門科目などといったように, 科目別の GPA も用いている. 打刻データとは, 先に述べた IC カード出欠管理システムにより蓄えられた学生の打刻のデータであり, このデータは「何年何月何日何時何分何秒に誰が打刻したか」をすべて記録ものである. 過去の研究 [4][6] において, この打刻データが成績予測や「要注意学生」の発見・予測に有用であることが証明されている. 打刻データから, 自動的に出欠データが生成される. 授業の開始時刻と終了時刻のそれぞれの前後の一定範囲内に打刻がある場合に有効打刻として出欠自動判定に用いられる. 出欠データには出欠の自動判定結果と判定に使用された授業開始時有効打刻時間と終了時有効打刻時間が記録されている. ところが, 授業の終了が早まったり, 遅れたりすることで打刻有効範囲がずれてしまい, 有効打刻として判定されない場合があった. また, 本来, 有効打刻として判定されるべき打刻が有効となっていなかった. これらの理由からこれまでは出欠データの信頼性からデータとして用いることができなかった. そこで本研究では出欠データを打刻データにより補正を行い, より正確なデータとして用いた. 打刻データでは誰がいつ打刻したかを記録しただけのものであるので, 授業に出席したのか, 早退したのか, 欠席したのか, ただの打刻し忘れなのか, それとも講義が休みだったのかを把握することができなかった. そのせいで欠席回数のような学生の授業に対する姿勢を如実に表す因子を厳密に調べることができなかった. 対して補正した出欠データで

は, 同じ授業をとっている学生同士を比べることにより, その日に授業があるのかないのか, また何時から何時まで授業があったのかを把握することができる. それゆえに, 欠席回数をより正確に数えることが可能になった. 本研究で定義した要注意学生において, 打刻データの代わりに出欠データを用いたことにより, 要注意学生の予測・発見に有用であることがわかった.

本論文では, 第 2 章において本研究で用いる手法の理論を述べ, 第 3 章では本研究に用いるデータの形式や拡張・変換・補正の方法を説明する. 第 4 章では,3 章で述べたデータを用いてベイジアンネットワークによる「要注意学生」の発見・予測とその検証を行った. そして第 5 章では本研究のまとめを述べる. ちなみに本研究では, 学生のデータを扱うにおいて, 個人を特定できる情報（指名や学籍番号）を一切排除した上で研究に着手しており, 本文によって個人情報が侵害されることはないことをここに付記する.

第2章　本研究に用いる手法の理論

本研究では分析及び予測の手法を多く用いている. その手法の多くはデータマイニングの知識発見の手法と同様である. 本章では属性選択, クラスタリング, ベイジアンネットワークについて説明する.

2.1　属性選択

属性選択は, 複数あるデータの中から有用なものを選択または合成することである.[7] 情報量が多すぎるとデータマイニングの有用性が失われてしまうことがある. 無関係な属性はデータにノイズをもたらし, 良い結果が得られない場合が多々ある. そこで属性選択を行い, データを取捨選択または合成することで, 結果を向上させることが期待できる. 本節では主成分分析と属性選択手法 Correlation based Feature Selection（以下 CFS）について説明する.

2.1.1　主成分分析

主成分分析とは, 複数の変数を持つデータの特徴を合成させて, 新たな変数を作り出す手法である. 今変数 $x_1, x_2, \ldots, x_n$ が存在するとして, 新たな変数 z_1 を導出するとした場合以下の式 2.1 ように表される。

$$z_1 = a_1x_1 + a_2x_2 + \ldots + a_nx_n \tag{2.1}$$

このとき, 各係数をベクトルとした $a_1, a_2, \ldots, a_n$ を,z_1 の分散が最大となるように各値を変化させる. だだしベクトル a の大きさが 1 となるという条件を満たす必要がある.

$$\sum_{i=1}^{n} a_i^2 = 1 \tag{2.2}$$

最大の分散が得られたとき, この z_1 を第 1 主成分とする. 次に第 1 主成分のときと同様に z_2 を以下の式 2.3 のように定める.

$$z_2 = b_1x_1 + b_2x_2 + \ldots + b_nx_n \tag{2.3}$$

このとき各係数をベクトルとした $b_1, b_2, \ldots, b_n$ を z_2 の分散が z_1 の分散の次に最大となるように各値を変化させる. ただしベクトル b の大きさが 1 となり, かつベクトル a とベクトル

b が垂直となるという条件を満たす必要ある.

$$\sum_{i=1}^{n} b_i^2 = 1 \tag{2.4}$$

$$\sum_{i=1}^{n} a_i b_i = 0 \tag{2.5}$$

こうして得られた z_2 を第 2 主成分とする. この作業を繰り返し行い, 主成分を作成する. この作業により多数の主成分が作成されるが, すべての主成分を用いることはせず, 十分にデータを説明することができる分だけを用いる. ではどのようにして主成分の数を決定するかというと, 寄与率と累積寄与率によって決定する. 寄与率とは, ある主成分が全体のデータの何 %を説明しているかを表している. ある主成分の固有値を λ_α, 各変数の分散を σ_i としたときの寄与率 C_α は以下の式 2.6 で求められる.

$$C_\alpha = \frac{\lambda_\alpha}{\sum_{i=1}^{n} \sigma_i} \tag{2.6}$$

また, 累積寄与率 P は寄与率の足すことで求められる.

$$P = \sum_{i=1}^{n} C_i \tag{2.7}$$

一般的に累積寄与率が 60 %〜80 %になるまで主成分を選択する.

2.1.2　情報利得と CFS

本研究では多くの変数を定義している. しかし, 先にも述べたように情報量が多ければ多いほど良いというわけではなく, 不必要なデータはノイズとなり結果に悪い影響をもたらしてしまう. これを回避するために, たくさんの変数の中から必要な変数を選ばなければならない. そこで挙げられるのが情報利得による属性選択である. 情報利得とは,2 つの確率分布との距離と説明される. ここでの距離とはあくまで表現としての距離である. 情報利得は 2 つの確率分布 P と Q を用いて, 以下の式 2.8 で定義される.

$$D(P||Q) = \sum_{x} P(x) \log \frac{P(x)}{Q(x)} \tag{2.8}$$

また, 情報利得は分割前の平均情報量と分割後の平均情報量の差でもある. そのため情報利得が最大となる属性を順番に選択することで, 決定木を構築することができる.

また, 情報利得を用いた変数選択の指標として,CFS が挙げられる. ある変数と関係性が強い変数は高い相関を持っていて, なおかつ他の変数と低い相関を持つという考えに基づき, 変数が選択される.CFS は以下の式 2.9 で求められる.k は変数の個数, Z は目的変数を指す. この CFS を最大化するような変数 Y_i を選択する. ちなみに SU は情報量 H と情報利得 D で求

めることができる.

$$CFS = \frac{\sum_{i=1}^{k} SU(Y_i, Z)}{\sqrt{k + \sum_{i=1}^{k} \sum_{j \neq i, j=1}^{k} SU(Y_i, Y_j)}} \tag{2.9}$$

$$SU(Y, Z) = 2 * \frac{D(Y||Z)}{H(Y) + H(Z)} \tag{2.10}$$

2.2　クラスタリング

クラスタリングとは, あるデータ群を類似性または非類似性に基づいてグループ分けをする手法であり, 教師なし学習に分類される. クラスタリングは階層的と非階層的とに大別することができる. 階層的クラスタリングは, 各データを 1 つのクラスタとして類似しているクラスタを併合する, または類似していないクラスタを別のクラスタにすることで, グループ分けをする手法である. 通常は 1 つのクラスタになるまで併合を繰り返す. 非階層的クラスタリングは, データの分割の良さを表す関数を定義して, その関数を最適化するようなクラスタ分けを探索する手法である. 本節では階層的クラスタリングの例として Ward's Method（以下ウォード法）, 非階層的クラスタリングの例として K-means 法を解説する.

2.2.1　ウォード法

ウォード法は, あるクラスタを併合した後のクラスタの分散と, 併合する前のクラスタそれぞれの分散の和との差が最小になるクラスタ同士を併合する手法である.$\sigma(x)$ をクラスタ x 内のデータの分散としたとき, 以下の式 2.11 で表される.

$$E_{i,j} = \sigma(x_i \cup x_j) - (\sigma(x_i) + \sigma(x_j)) \tag{2.11}$$

この $E_{i,j}$ が最小になるように, クラスタを併合していく. またこの手法ははずれ値に強い性質を持っている.

2.2.2　k-means 法

k-means 法は k-平均法とも呼ばれ, 多数のデータをいくつかのクラスタに分類する手法である. 階層的クラスタリングとの違いは, クラスタ数を分類する前に設定しておかなければならないという点である. あるデータ群を k 個のクラスタに分類する場合, 次の手順で行われる.

1. k 個のデータをランダムで選択しシード値を生成する.
2. 別のデータ 1 つに対して, 最もシード値の近いクラスタ求め, データをそのクラスタに分類する.
3. 各クラスタのシード値を生成する.

2 と 3 の手順を繰り返し, すべてのデータの分類が終わるまで続ける. 手順 2 において, あるデータとクラスタとの距離を求める. その距離の指標はユークリッド距離が最も有名であり, 一般的である. k-means 法の利点は, 階層的クラスタリングよりも高速に実行することができ, 実装も用意であるという点である. しかし, 最初の k 個のデータはランダムで選択されるため, クラスタ数や初期のシード値をに大きく影響を受けてしまったり, 再現性に乏しいという欠点がある.

2.3　ベイジアンネットワーク

ベイジアンネットワーク [8][9] は事象同士の依存関係があると推論し, それを有効グラフで表した確率モデル（グラフィカルモデル）である. この特性を応用して, 不確実性を含む事象の予測や合理的な意思決定, 観測結果から原因を探る故障診断などに用いられている [10].

ベイジアンネットワークは確率変数, 有効グラフ構造, 条件付き確率で定義される. この 3 つの要素を決定することは, ベイジアンネットワークのモデルを作成することである. それゆえに, 最適なベイジアンネットワークのモデルを作成するには, 最適な条件付確率の推定, 最適な確率変数の選択, 最適な有効グラフの獲得が必要不可欠となる.

2.3.1　ベイジアンネットワークによる予測

図 2.1 はベイジアンネットワークの例である. 確率変数 X_1 と X_2 の間の依存関係を $X_1 \to X_2$ と表されている. この場合 X_1 を親ノード,X_2 を子ノードとして扱われる. 子ノード X_2 の親ノードを $P_a(X_2)$ とすると,X_2 と $P_a(X_2)$ の依存関係は $P(X_2|P_a(X_2))$ という条件付確率で表せる. 図 2.1 の 4 つの確率変数 X_1, X_2, X_3, X_4 について考えた場合, すべての確率変数の同時確率分布 $P(X_1 \ldots X_4)$ は以下の式 2.12 のように表せる.

$$P(X_1 \ldots X_4) = \prod_{i=1}^{4} P(X_i|P_a(X_i)) \tag{2.12}$$

すべての変数の事後確率は, 同時確率分布を計算することで求められるので, ベイジアンネットワークはこれを用いることで得ることができる. しかしこのように事後確率を計算すると, 変数が n 個あったとすると指数オーダーのサイズが必要となり n が大きくなると実用的ではなくなってしまう. そこで計算コストを削減するため, あるノードとその親ノードと子ノードに注目した局所的確率計算により事後確率を計算する. 観測された情報からの確率伝播（変数間の局所計算）によって確率分布を更新していくことから確率伝播法と呼ばれている. 図 2.2 の構造をもとでの計算の実行例を示す.

$X_1 \to X_2, X_2 \to X_3$ の間に依存関係があり, 条件付確率が与えられているとする. 計算しようとしているノードを X_2 として, 観測された変数の値を e とすると X_2 の事後確率は $P(X_2|e)$ と表せる. また,X_2 よりも上流に存在するノード群（親ノード群）に入力される観測情報と, X_2 よりも下流に存在するノード群（子ノード群）に入力される観測情報として

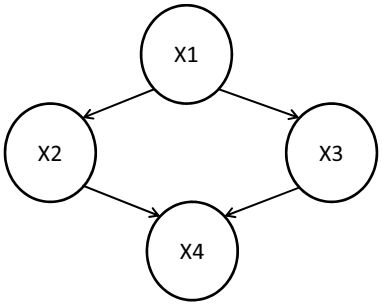

図 2.1: ベイジアンネットワークの例

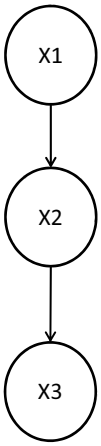

図 2.2: モデルの一部分

それぞれ e^+, e^- を与えとき, 事後確率 $P(X_2|e)$ はベイズの定理により以下の式 2.13 のように表せる.

$$\begin{aligned} P(X_2|e) &= P(X_2|e^+, e^-) \\ &= \frac{P(e^-|X_2, e^+)P(X_2|e^+)}{P(e^-|e^+)} \end{aligned} \tag{2.13}$$

e^+ と e^- は X_2 に依存しないものであるので, 定数 $\alpha = \frac{1}{P(e^-|e^+)}$ として扱うことで式 2.13 は次のように変形できる.

$$P(X_2|e) = \alpha P(e^-|X_2, e^+)P(X_2|e^+) \tag{2.14}$$

このうち親ノードから X_2 へ伝播する確率を $P(X_2|e^+) = \pi(X_2)$ とする. $\pi(X_2)$ はすでに定義している $P(X_2|X_1)$ と $P(X_1|e^+)$ によって計算が可能である.

$$\pi(X_2) = \sum_{X_1} P(X_2|X_1)P(X_1|e^+) \tag{2.15}$$

X_1 に親ノードがない場合は予め用意された事前確率を与え, 観測情報が与えられている場合, その値は決定できる. X_1 に入力がなく, かつ X_1 に親ノードが存在するとき式 (2.14) を再帰的に適用することによりその値を求めることができる. 子ノードから X_2 へ伝播する確率を $P(e^-|X_2) = \lambda(X_2)$ として, 式 2.15 と同様に考えると次のように表せる.

$$\lambda(X_2) = \sum_{X_3} P(X_3|X_2)P(e^-|X_2, X_3) \tag{2.16}$$

観測から得られた情報 e^- は X_2 の値に関係なく独立であることから

$$\lambda(X_2) = \sum_{X_3} P(X_3|X_2)P(e^-|X_3) \tag{2.17}$$

とすることができる.$P(X_3|X_2)$ はすでに定義されていることから, 観測情報が与えられているとき値が決定できる. また, 観測情報がなく X_3 が子ノードを持たない下端のノードの場合は, 無情報であることから一様分布確率として X_3 のあらゆる状態について等しい値とする. また,X_3 が子ノードを持つ場合,$\pi(X_2)$ の場合と同様に, 式 2.16 を再帰的に適用することで最終的に下端のノードの値を求めることができる. このようにして X_2 の事後確率を確率伝播法によって局所的に求めることで, 計算コストを削減することができる.

しかし確率伝播法はどのようなグラフ構造でも厳密な値を算出できるとは限らない. ベイジアンネットワークを無効グラフとした場合, ノードとノードを繋ぐパス全てがループを持たない時, そのベイジアンネットワークは singly connected と呼び, パスがどこか 1 か所でもループを持つ時, multiply connected と呼ぶ. グラフ構造が singly connected であるならば, 上端のノードと下端のノードが求めることができるので, 確率伝播法によって厳密な値を算出することができる. しかし, グラフ構造が multiply connected である時, ループを持っているため, 上端のノードと下端のノードを求めることができない場合がある. その場合, 単純に確率を伝播していくだけでは, 計算を収束させることができない可能性がある. そこでグラフ構造を multiply connected なグラフと同等な singly connected なグラフに変換し, その上で確率伝播法を適用する手法である. この手法を Junction Tree アルゴリズムと呼ぶ. このアルゴリズムが開発されたことにより, ベイジアンネットワークに対する有用性が高まり, 技術発展やシステム開発が方々で進められている.

2.3.2　確率変数

ベイジアンネットワークに用いられる確率変数は, 離散値であることが望ましい. つまり数値変数は離散化する必要がある. たとえば, あるテストの点数があったとして, その点数は数値で記録されているので, そのままベイジアンネットワークに適用はしない.

離散化の手法として, データを等分割するか, クラスタリングによる分割を行う. データの等分割は, データが 100 個あるとすると,33 個,33 個,34 個というように分割する. クラスタリングによる分割は, データをクラスタリングによって分割することで, 分割されてできた集合 1 つ 1 つに意味を持たせることができる.

2.3.3　有効グラフ構造

ベイジアンネットワークのモデルは有効グラフで表されている. よって有効グラフの構造がベイジアンネットワークの予測結果に大きく影響する. ベイジアンネットワークの有効グラフ構造にはいくつか種類があり, その代表的な構造について簡単に説明する.

Naive Bayes

Naive Bayes は図 2.3 のように目的変数を上端の親ノードつまり木構造における根の部分に置き, 残りの変数をすべて根ノードの葉としたものである. 目的変数の事後確率はベイズの定理により求められ, グラフの構造もベイジアンネットワークにおいて最もシンプルな構

造をしている. それゆえ, 実装が簡単で学習時間が短いという利点がある. しかし, 葉となるノードが多ければ予測精度が向上するわけではなく, むしろ下がる可能性すらある. 説明変数の選択には注意が必要である.

Tree Augmented Network

Tree Augmented Network（以下 TAN）は, 図 2.4 のように Naive Bayes 構造の子ノードから, 目的変数以外にもう一つだけ親ノードとして持っている構造をしている. 親ノードの選択基準として条件付相互情報量が用いられている. ある確率変数 X,Y として目的変数 C が与えられる条件付相互情報量は

$$I(X,Y|C) = -\sum_{X}\sum_{Y}\sum_{C} P(x,y,c)\log\frac{P(x,y|c)}{P(x|c)P(y|c)} \tag{2.18}$$

と表せる.

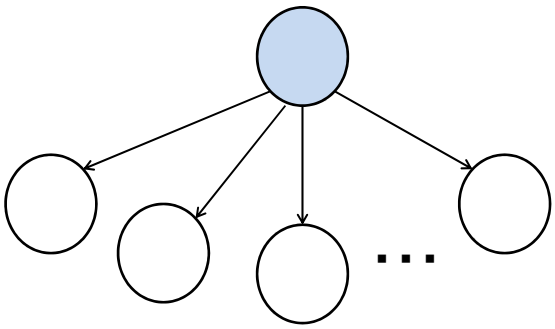

図 2.3: Naive Bayes の例

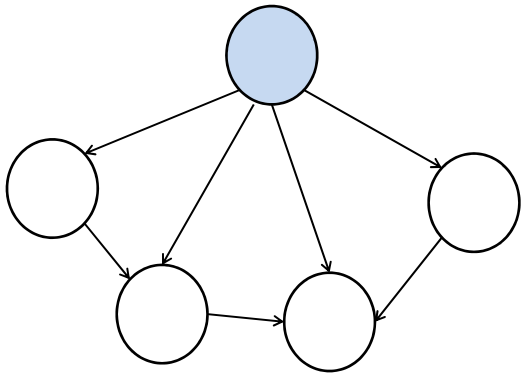

図 2.4: TAN の例

Free Network

Free Network は親ノードと子ノードの数に制限を設けていないグラフ構造である. しかしあるノードに対する親ノードの数が増えていくと, 条件付確率は大幅に増えていく. そのため親ノードの数を制限してグラフの構築をする場合が多い.

第3章　本研究に用いるデータの概要とその拡張及び変換

本研究では「要注意学生」の発見・予測の手法としてベイジアンネットワークを採用している. ベイジアンネットワークによる予測は, データの質によって良し悪しが決まるといっても過言ではない. 本章では用いたデータの概要とその拡張及び変換について解説をしていく.

3.1　用いるデータの概要

本研究では, 名古屋工業大学を在籍していた 338 名の学生に関するデータを用いている. この 338 名はある 2 つの年度の学生たちであり, それぞれ 171 名と 167 名である. データは 4 種類あり, 講義別成績データ, 誰がいつ打刻をしたかを記録したデータ（以下打刻データ）, 誰がいつどの授業の出席または欠席したかを記録したデータ（以下出欠データ）, そして学生が卒業研究に着手した年次や卒業した年次, 退学した年次, 退学した理由が記載されたデータ（以下修学データ）である.

3.1.1　講義別成績データ

学生の講義別成績データは, 学籍番号, 講義の成績, 授業名, 授業が開かれた年次と時期の 4 つの情報をレコード形式で保存されている. ちなみに, 学籍番号は暗号化されており個人を特定できないようになっている. また, 記載されている授業名は実際の授業名ではなく,「専門 1」や「演習 1」のように講義を特定できないようされている. これは学生の学科を特定し, 個人を推測されないようにする措置であり, 英語や理系基礎科目, リベラルアーツなどのすべての学科の学生が受ける授業の名前は変更されていない. そのため, 具体的な講義の内容はわからないが, 講義の分野は知ることができる.

講義の成績は, 秀・優・良・可・不可・失格の 6 つの評価がある. 秀が最もよい成績で, 秀・優・良・可・不可と成績の評価が悪くなっていく. 成績が秀・優・良・可であるならば単位取得が認められ, 不可・失格であれば認められない. 不可と失格の違いは, 課題提出やテストを受験していながら単位取得の条件を満たすことができなかった場合は成績が不可となり, 課題未提出やテストを受験できなかった場合, 出席回数が既定の回数を満たすことができなかった場合は, 成績の評価ができないとして失格となる.

3.1.2　打刻データ

打刻データは, 第 1 章でも述べたように,IC カード出欠管理システムにより蓄えられた学生の打刻のデータであり, このデータは「何年何月何日何時何分何秒に誰が打刻したか」をすべて記録ものである. 実際にはレコード形式で保存されており, 学籍番号, 打刻した日付（年/月/日）, 打刻した時間の 3 つで構成されている. 講義別成績データと同じく学籍番号は暗号化されているが, 講義別成績データの学籍番号と共通であるため, 打刻データと講義別成績データを関連付けることは簡単である.

3.1.3　出欠データ

打刻データと同様にレコード形式で保存されており, 授業番号, 学籍番号, 自動出欠判定, 入室打刻をした日付・時間, 退室打刻をした日付・時間である. 学籍番号は暗号化されており, これまでと同様に学籍番号は共通している. 出欠データは打刻データから自動的に生成される. 授業の開始時刻と終了時刻のそれぞれの前後の一定範囲内に打刻がある場合に有効打刻として出欠自動判定に用いられる. 出欠データには出欠の自動判定結果と判定に使用された授業開始時有効打刻時間と終了時有効打刻時間が記録されている. ところが, 授業の終了が早まったり, 遅れたりすることで打刻有効範囲がずれてしまい, 有効打刻として判定されない場合があった. また, 本来, 有効打刻として判定されるべき打刻が有効となっていなかった. この点を考慮した上でこのデータを扱う必要がある.

3.1.4　修学データ

修学データは, 本研究対象である 338 名の学生の卒業研究に着手した年次や卒業までにかかった年数, 退学をした学生の退学年次, 退学した学生の退学理由が記載されたデータである. 名古屋工業大学では, 4 年生から研究室に入り卒業研究に着手することになっている. 卒業研究に着手するには条件があり, 所定の単位数を取得していなければならない.3 年生の終了時に条件を満たしてしなければ, 実質的に留年となる. また, 卒業するためにも条件があり, 卒業研究に着手する条件と同じように所定の単位数が必要となる. 4 年次に卒業研究に着手することができても, その年度に卒業できない場合もありうる.

3.2　データの拡張及び変換

講義別成績データや打刻データ, 出欠データはレコード形式であり, そのデータすべての量は 70 万にも達する. しかしながらベイジアンネットワークからモデルを構築するにはある程度の情報量が必要となる. これらのデータはデータの数は多くとも情報量には乏しい. ゆえにこのままの形式で用いても, 満足のいくモデルが構築することはできない. そこで本研究ではデータの拡張及び変換を行った.

3.2.1　講義別成績データの拡張及び変換

本研究では, 成績の指標として GPA を用いる.GPA は各成績の評価である秀・優・良・可・不可・失格にそれぞれ 4 点・3 点・2 点・1 点・0 点・0 点と得点を割り振り, 講義毎に決められている単位数を用いて式 3.1 により求められる.

$$GPA = \frac{\sum^{\text{受講した講義全て}}(\text{成績得点}) * (\text{講義の単位数})}{\sum^{\text{受講した講義全て}}(\text{講義の単位数})} \tag{3.1}$$

本研究では, 全学生の各年次別の年間・前期・後期の GPA の他に, 講義を分野毎に分類し, 分野別の GPA も求めた. また各成績の評価の各年次別の前期・後期の獲得数も求めた. GPA が 2．0 であっても, すべての教科が良である場合と, 秀と不可の極端な成績の場合が考えられる. 以下の表 3.1 にここで述べた変数を示す.

3.2.2　打刻データの拡張及び変換

打刻データに関しては打刻の回数に着目した. 打刻した日付から月別打刻回数を求めた. 以下の表 3.2 にここで述べた変数示す.

3.2.3　出欠データの補正及び拡張

打刻データから, 自動的に出欠データが生成される. 授業の開始時刻と終了時刻のそれぞれの前後の一定範囲内に打刻がある場合に有効打刻として出欠自動判定に用いられる. 出欠データには出欠の自動判定結果と判定に使用された授業開始時有効打刻時間と終了時有効打刻時間が記録されている. ところが, 授業の終了が早まったり, 遅れたりすることで打刻有効範囲がずれてしまい, 有効打刻として判定されない場合があった. また, 本来, 有効打刻として判定されるべき打刻が有効となっていなかった. そこで本研究では, 打刻データを用いて出欠データの補正を行った. 打刻データでは記録があっても出欠データでは記録されていない打刻が存在する. この打刻を両者の打刻記録を比較し, 記録されていない打刻を補完する. また, 打刻データで記録されていない打刻の補完を行ったとしても, それでも打刻の記録が存在しない箇所がある. その打刻は入退室の打刻を忘れてしまった場合や IC カードリーダーの不具合, そもそも体育や一部の実験室などの IC カードリーダーが設置されていない教室での授業だと考えられる. そのため入退室の打刻をしていないからと言って, 安易に遅刻・早退・欠席とすることはできない. しかし出欠データには授業番号が記載されている. 同じ授業を受講している学生を比較することで, 授業が休講であったのか,IC カードリーダーが設置されていない教室での授業だったのか, それとも欠席であったのかを判断することができる.

こうして補正を行ったデータを拡張・変換を行う. まずは出席の回数を求める. ここでは入室と退室のどちらかを打刻している回数を出席の回数としてデータを作成した. 理由は, 先に述べたように入退室の打刻を忘れてしまった場合や IC カードリーダーの不具合が考えられるからである. さらに欠席の回数を求めた. ここでは入室と退室の打刻を両方とも行っていない回数を求めた. 先ほどの補正のおかげで欠席はかなり正確だと考えられるからである.

欠席回数は, 通年欠席回数と前期欠席回数, 後期欠席回数を求めた. 以下の表 3.3 にここで述べた変数を示す.

表 3.1: 成績に関する変数

番号	変数名	意味
1	1 年次通年	1 年次に受講した講義の GPA
2	1 年次前期	1 年次の前期に受講した講義の GPA
3	1 年次後期	1 年次の後期に受講した講義の GPA
4	外国語 1 年次前期	1 年次前期に受講した外国語に関係する講義の GPA
5	外国語 1 年次後期	1 年次後期に受講した外国語に関係する講義の GPA
6	人文 1 年次前期	1 年次前期に受講した人間文化に関係する講義の GPA
7	人文 1 年次後期	1 年次後期に受講した人間文化に関係する講義の GPA
8	数学 1 年次前期	1 年次前期に受講した数学に関係する講義の GPA
9	数学 1 年次後期	1 年次後期に受講した数学に関係する講義の GPA
10	理科 1 年次前期	1 年次前期に受講した理科に関係する講義の GPA
11	理科 1 年次後期	1 年次後期に受講した理科に関係する講義の GPA
12	体育 1 年次前期	1 年次前期に受講した体育に関係する講義の GPA
13	体育 1 年次後期	1 年次後期に受講した体育に関係する講義の GPA
14	専門 1 年次前期	1 年次前期に受講した専門科目に関係する講義の GPA
15	専門 1 年次後期	1 年次後期に受講した専門科目に関係する講義の GPA
16	その他 1 年次前期	1 年次前期に受講した上記の講義に分類されない講義の GPA
17	その他 1 年次後期	1 年次後期に受講した上記の講義に分類されない講義の GPA
18	前期秀獲得数	1 年次の前期に秀を獲得した数
19	後期秀獲得数	1 年次の後期に秀を獲得した数
20	前期優獲得数	1 年次の前期に優を獲得した数
21	後期優獲得数	1 年次の後期に優を獲得した数
22	前期良獲得数	1 年次の前期に良を獲得した数
23	後期良獲得数	1 年次の後期に良を獲得した数
24	前期可獲得数	1 年次の前期に可を獲得した数
25	後期可獲得数	1 年次の後期に可を獲得した数
26	前期不可獲得数	1 年次の前期に不可を獲得した数
27	後期不可獲得数	1 年次の後期に不可を獲得した数
28	前期失格獲得数	1 年次の前期に失格を獲得した数
29	後期失格獲得数	1 年次の後期に失格を獲得した数

表 3.2: 打刻に関する変数

番号	変数名	意味
30	1 年次 4 月打刻回数	1 年次の 4 月に打刻した回数
31	1 年次 5 月打刻回数	1 年次の 5 月に打刻した回数
32	1 年次 6 月打刻回数	1 年次の 6 月に打刻した回数
33	1 年次 7 月打刻回数	1 年次の 7 月に打刻した回数
34	1 年次 8 月打刻回数	1 年次の 8 月に打刻した回数
35	1 年次 9 月打刻回数	1 年次の 9 月に打刻した回数
36	1 年次 10 月打刻回数	1 年次の 10 月に打刻した回数
37	1 年次 11 月打刻回数	1 年次の 11 月に打刻した回数
38	1 年次 12 月打刻回数	1 年次の 12 月に打刻した回数
39	1 年次 1 月打刻回数	1 年次の 1 月に打刻した回数
40	1 年次 2 月打刻回数	1 年次の 2 月に打刻した回数
41	1 年次 3 月打刻回数	1 年次の 3 月に打刻した回数

表 3.3: 出欠に関する変数

番号	変数名	意味
42	1 年次 4 月出席回数	1 年次の 4 月に入室または退室の打刻をした回数
43	1 年次 5 月出席回数	1 年次の 5 月に入室または退室の打刻をした回数
44	1 年次 6 月出席回数	1 年次の 6 月に入室または退室の打刻をした回数
45	1 年次 7 月出席回数	1 年次の 7 月に入室または退室の打刻をした回数
46	1 年次 8 月出席回数	1 年次の 8 月に入室または退室の打刻をした回数
47	1 年次 9 月出席回数	1 年次の 9 月に入室または退室の打刻をした回数
48	1 年次 10 月出席回数	1 年次の 10 月に入室または退室の打刻をした回数
49	1 年次 11 月出席回数	1 年次の 11 月に入室または退室の打刻をした回数
50	1 年次 12 月出席回数	1 年次の 12 月に入室または退室の打刻をした回数
51	1 年次 1 月出席回数	1 年次の 1 月に入室または退室の打刻をした回数
52	1 年次 2 月出席回数	1 年次の 2 月に入室または退室の打刻をした回数
53	1 年次 3 月出席回数	1 年次の 3 月に入室または退室の打刻をした回数
54	1 年次前期欠席回数	1 年次前期に入室と退室両方の打刻がされていない回数
55	1 年次後期欠席回数	1 年次後期に入室と退室両方の打刻がされていない回数
56	1 年次通年欠席回数	1 年次に入室と退室両方の打刻がされていない回数

第4章　要注意学生の発見

この章では要注意学生の発見手法を提案する. そのためにまず要注意学生を厳密に定義する. 厳密に定義することで, 要注意学生の予測・発見をより良くすることを期待する. 本研究で用いる変数は先にも記述した通り 56 個ある. これらをすべて使っても, 良い結果が得られるわけではない. ゆえに変数を取捨選択して発見に用いる. さらにこれらの変数は数値で表されているので, これらを離散化する必要がある. これらの下準備を行ってから要注意学生の発見を行う.

打刻データと成績データを用いた要注意学生の発見や成績予測の有用性について他の研究で示されている. 本研究では, 要注意学生を発見するのに打刻データを出欠データに見直している. そこで打刻データと成績データを用いた要注意学生の発見精度と, 打刻データを出欠データに変更した要注意学生の発見精度比較することで本提案の有用性を検証する. また, 本研究での要注意学生の定義と従来の定義との比較検証も行う.

4.1　発見の下準備

要注意学生の発見に際して, 予測して発見されるのはいったい誰なのか, どのタイミングで発見を行うのか, そしてその発見がいったいどれぐらい有用性があるのかの評価はどうするのかを決定しなければならない. また, 先述したように, 変数を取捨選択して離散化する必要がある. これらを以下により述べる.

4.1.1　発見を行う時期

学生が入学して幾日と立たない間に要注意学生の予測を行おうとしても, 情報が少なく有用な発見は不可能に近い. また, 時期を遅らせ, 情報を多く得ようとしても, 指導が遅く手遅れとなる可能性が高い. 情報の量と指導の効果のバランスを考える必要がある. そこで本研究では 1 年次終了した時点での指導を考える. つまり,1 年次の前期と後期の情報を使い,2 年次以降の修学状況を予測する.

4.1.2　発見の対象者と要注意学生

本研究における発見の理想は,「一見優秀であるが将来的に修学状況が悪化する」学生を予測し発見することである. たとえば,GPA の値が 0 に近いような学生は, 指導が必要であるということは誰にでも判断することができるが, GAP の値が平均的である学生が指導を必要としているかどうかは判断しづらい. 後者の学生をみつけることを本研究では目指す.

表 4.1: 1 年前期・後期の GPA 別人数

	1 年前期 GPA			1 年後期 GPA		
値域	全人数	留年・退学	割合	全人数	留年・退学	割合
0 以上 0.5 未満	5	5	100 %	13	13	100 %
0.5 以上 1.0 未満	8	6	75 %	12	12	100 %
1.0 以上 1.5 未満	11	7	64 %	31	14	45 %
1.5 以上 2.0 未満	48	23	48 %	67	15	22 %
2.0 以上 2.5 未満	105	16	15 %	96	10	10 %
2.5 以上 3.0 未満	100	10	10 %	70	2	3 %
3.0 以上 3.5 未満	53	3	6 %	42	4	10 %
3.5 以上 4.0 未満	8	0	0 %	7	0	0 %
	338	70		338	70	

では具体的にはどうであるか. 表 4.1 に 1 年次の前期と後期の GPA 別に, 含まれている学生の人数と, 留年者または退学者の人数を示す.

表 4.1 を見ると留年または退学している学生は 70 名存在していて, 全体の約 21 %である. また,GPA の値が高くなるにつれて, 留年または退学している学生の割合は減っているのがわかる. さらに,1 年次の前期または後期の GPA が 1.0 を下回っている学生のほとんどが留年または退学している. 1 年次後期に注目してみると,GPA1.0 未満の学生は 25 名中 25 名が留年または退学をしている. すなわち,1 年次の GPA が前期か後期のどちらか一方でも 1.0 を下回れば, 留年または退学する可能性がかなり高いと考えることができる. そこで本研究では,1 年次の GPA が前期か後期のどちらか一方でも 1.0 を下回っている学生は, 指導が必要であることは明白であるため, 必然的に指導対象者とする. そして発見の対象は 1 年次 GPA が 1.0 を上回っている学生とし, 要注意学生として発見されるべき学生は,「1 年次 GPA が 1.0 を上回っていて, 将来的に留年または退学する学生」とする. これらの学生の具体的人数は, 「1 年次の GPA が前期か後期のどちらか一方でも 1.0 を下回っている学生」とした指導対象者は 31 人, 「1 年次 GPA が 1.0 を上回っている学生」とした発見対象者は 307 人, 発見対象者の中の「1 年次 GPA が 1.0 を上回っていて, 将来的に留年または退学する学生」とした要注意学生は 41 人となっている. また, これは従来の研究の定義である.

しかしこの要注意学生の定義だと, 大学院入試や就職活動の失敗によって計画的に留年した学生や, 転学のために退学した学生など教員によるアドバイスや指導をあまり必要と考えていない学生まで含まれてしまう. 本来発見したい学生は, 学業の不信や大学に馴染むことができない学生たちであり, そういった学生こそアドバイスや指導を必要としている. そこで発見対象者と要注意学生をさらに調査することにした. まず発見対象者 307 名（A 年度 155 名,B 年度 152 名）の卒業研究に着手した年数と卒業にかかった年数を調べた. 用いたデータは第 3 章で説明した修学状況が記載されている修学データである.

名古屋工業大学では 4 年次から卒業研究に着手する. しかし 3 年次終了時に, 指定の単位取得条件を満たしていない場合は卒業研究に着手することができず, 事実上留年となってしまう. 卒業研究に着手した年数を以下の表 4.2 に示す. さらに卒業にかかった年数を以下の

表 4.3 に示す.

表 4.2: 各年度における卒業研究着手の年数の分布

	3 年	4 年	5 年	6 年	未着手	退学	合計
A 年度	141	7	1	2	2	2	155
B 年度	137	9	1	0	2	3	152

表 4.3: 各年度における卒業にかかった年数の分布

	4 年	5 年	6 年	在学	退学	合計
A 年度	133	14	1	4	3	155
B 年度	133	9	0	4	6	152

この表 4.2 において「未着手」とは記録上卒業研究に着手をしていないことを意味している.「退学」とは卒業研究に着手する前に退学届が提出されていることを意味している. 表 4.3 において「在学」とは卒業や退学をしていな状態を意味している.「退学」とは卒業までに退学届が提出されていることを意味している. 表 4.2 を見ると卒業研究に 4 年次に着手できたのは 278 名いるが, 表 4.3 を見ると 4 年間で卒業できたのは 266 名と 12 名が順調に卒業していないことがわかる. そこで卒業研究を 4 年次に着手できた学生の卒業にかかった年次を以下の表 4.4 に示す.

表 4.4: 卒業研究を 4 年次に着手できた学生の卒業にかかった年数の分布

	4 年	5 年	6 年	在学	退学	合計
A 年度	133	7	0	0	1	141
B 年度	133	0	0	1	3	137

表 4.4 を見ると 5 年間で卒業が 7 名, まだ在学しているのが 1 名, そして順調に 4 年次まで進級できたのにもかかわらず退学してしまっている学生が 4 名もいる. 5 年間で卒業やまだ在学中の学生は, 大学院入試や就職活動の失敗などの計画的が考えられる. 順調に 4 年次まで新旧できたのにもかかわらず退学してしまっている学生は, なぜ退学してしまったのだろうか.307 名の発見対象者の中には 9 名の退学者がいる. この 9 名の退学者全員の退学理由と合わせて調査していく.

表 4.5: 在学年数と退学理由と卒研着手年数

	在学年数	理由	着手年数
1	0 年間	転学科	未着手
2	2 年間	他大学受験	未着手
3	2 年間	他大学受験	未着手
4	3 年間	不明	未着手
5	1 年間	他大学受験	未着手
6	4 年間	就職	3 年
7	5 年間	授業料未納	未着手
8	4 年間	就職	3 年
9	4 年間	一身上の都合	3 年

表 4.5 に退学者の退学理由とその者の在学年数, 卒業研究に着手した年数を示す. 上から 5 名は 3 年以内に大学を退学した者, その下の 4 名は 4 年以上大学に在籍した者である. 3 年以内に大学を退学した者のほとんどは, 他大学への受験を行い退学している. 4 年以上大学にいる者の半分は就職となっている. さらに 1 名を除いて,3 年で卒業研究に着手している. こうして退学者の調査をしてみると, 真に指導が必要だといえる者は, 4 年以上在学しており,3 年間で卒業研究に着手していない学生だと考えられる.

これらの調査結果をふまえて要注意学生を次のように定義した.

要注意学生の定義

- GPA が 1.0 以上で留年または退学してしまう学生
- 3 年間で卒業研究に着手した者は要注意学生から外す
- 入学から 3 年以内に退学した者はデータから除外する

このように要注意学生を定義することで, より指導やアドバイスが必要な学生を予測し発見することが期待できる. また, この定義より要注意学生は 41 名から 25 名に, 予測対象者は 307 名から 302 名になった.

4.1.3　変数選択

本研究で用いるデータは第 3 章でも述べたように, 成績データ, 打刻データ, 出欠データの 3 種類である. 成績データは GPA 値と獲得成績数にデータを拡張し変数の数は 29 個, 打刻データは月別の打刻回数と前期・後期・通年の打刻回数に拡張し, 変数の数は 12 個, 出欠データは月別の出席回数と前期・後期・通年の出席回数, 欠席回数にデータを拡張し, 変数の数は 15 個となり, 合計 56 個となっている. この変数をすべて同時に用いても良い結果が出る可能性は低い. そのためデータを組み合わせたり, 一部削除することにする. また, 属性変数の取

捨選択には, 属性選択の手法として用いられる CFS を採用した. 本研究では出欠データを用いた発見の有用性を示すため打刻データを用いた発見との比較を行う. それを考慮してデータを組み合わせる. 変数の組み合わせ方は以下の表 4.6, 表 4.7 に示す. ちなみに 9 月と 3 月は大学が長期休暇に入っていて, 授業が開講されていないため, ほとんどの学生が打刻や出席が 0 回である. そのため 9 月と 3 月の打刻回数と出席回数は削除した.

これらの変数を用いて CFS による属性選択を行い, より要注意学生の特徴を表している変数を選択する. この選択は, 本研究の要注意学生の定義と従来の定義それぞれに行う. まず従来の定義の場合の CFS による属性選択を行ったところ, 打刻データを用いた変数の組み合わせでは, 36 変数の中から 14 変数が選択された. また, 出欠データを用いた変数の組み合わせでは, 39 変数の中から 13 変数が選択された. 選択された変数をそれぞれ表 4.8 と表 4.9 に示す. 次に本研究の定義の場合の CFS による属性選択を行ったところ, 打刻データを用いた変数の組み合わせでは, 36 変数の中から 11 変数が選択された. また, 出欠データを用いた変数の組み合わせでは, 39 変数の中から 13 変数が選択された. 選択された変数をそれぞれ表 4.10 と表 4.11 に示す.

表 4.6: 打刻データと成績データの変数組み合わせ

番号	変数群
1	外国語 1 年次前期
2	外国語 1 年次後期
3	人文 1 年次前期
4	人文 1 年次後期
5	数学 1 年次前期
6	数学 1 年次後期
7	理科 1 年次前期
8	理科 1 年次後期
9	体育 1 年次前期
10	体育 1 年次後期
11	専門 1 年次前期
12	専門 1 年次後期
13	その他 1 年次前期
14	その他 1 年次後期
15	前期秀獲得数
16	後期秀獲得数
17	前期優獲得数
18	後期優獲得数
19	前期良獲得数
20	後期良獲得数
21	前期可獲得数
22	後期可獲得数

23	前期不可獲得数
24	後期不可獲得数
25	前期失格獲得数
26	後期失格獲得数
27	1 年次 4 月打刻回数
28	1 年次 5 月打刻回数
29	1 年次 6 月打刻回数
30	1 年次 7 月打刻回数
31	1 年次 8 月打刻回数
32	1 年次 10 月打刻回数
33	1 年次 11 月打刻回数
34	1 年次 12 月打刻回数
35	1 年次 1 月打刻回数
36	1 年次 2 月打刻回数

表 4.7: 出欠データと成績データの変数組み合わせ

番号	変数群
1	外国語 1 年次前期
2	外国語 1 年次後期
3	人文 1 年次前期
4	人文 1 年次後期
5	数学 1 年次前期
6	数学 1 年次後期
7	理科 1 年次前期
8	理科 1 年次後期
9	体育 1 年次前期
10	体育 1 年次後期
11	専門 1 年次前期
12	専門 1 年次後期
13	その他 1 年次前期
14	その他 1 年次後期
15	前期秀獲得数
16	後期秀獲得数
17	前期優獲得数
18	後期優獲得数

19	前期良獲得数
20	後期良獲得数
21	前期可獲得数
22	後期可獲得数
23	前期不可獲得数
24	後期不可獲得数
25	前期失格獲得数
26	後期失格獲得数
27	1 年次 4 月出席回数
28	1 年次 5 月出席回数
29	1 年次 6 月出席回数
30	1 年次 7 月出席回数
31	1 年次 8 月出席回数
32	1 年次 10 月出席回数
33	1 年次 11 月出席回数
34	1 年次 12 月出席回数
35	1 年次 1 月出席回数
36	1 年次 2 月出席回数
37	1 年次通年欠席回数
38	1 年次前期欠席回数
39	1 年次後期欠席回数

4.1.4　変数の離散化

ベイジアンネットワークに用いる確率変数は, 離散値であることが望ましい. 本研究で用いている確率変数は, 数値で表されているので離散化する必要がある. ゆえに先の変数選択で選択された変数を離散化する. 本研究では離散化の手法として, ウォード法によるクラスタリングを用いる. 離散化する際の属性数は全変数において 3 あるいは 4 とした. 離散化した結果を以下に示す. 表 4.8 に示した打刻データと成績データの変数群の結果は表 4.12 と表 4.13, 表 4.9 に示した出欠データと成績データの変数群の結果は表 4.14 と表 4.15 である. また, 表 4.10 に示した打刻データと成績データの変数群の結果は表 4.16 と表 4.12, 表 4.11 に示した出欠データと成績データの変数群の結果は表 4.18 と表 4.19 である.

表 4.8: 従来の定義で選択された変数の組み合わせ（打刻データ）

番号	打刻データを用いた変数群
1	人文 1 年次後期
2	数学 1 年次後期
3	理科 1 年次前期
4	理科 1 年次後期
5	専門 1 年次後期
6	前期不可獲得数
7	後期秀獲得数
8	後期不可獲得数
9	後期失格獲得数
10	1 年次 5 月打刻回数
11	1 年次 6 月打刻回数
12	1 年次 7 月打刻回数
13	1 年次 12 月打刻回数
14	1 年次 1 月打刻回数

表 4.9: 従来の定義で選択された変数の組み合わせ（出欠データ）

番号	出欠データを用いた変数群
1	人文 1 年次後期
2	数学 1 年次後期
3	理科 1 年次後期
4	専門 1 年次後期
5	前期不可獲得数
6	後期秀獲得数
7	後期不可獲得数
8	後期失格獲得数
9	1 年次 7 月出席回数
10	1 年次 12 月出席回数
11	1 年次 1 月出席回数
12	1 年次 2 月出席回数
13	1 年次前期欠席回数

表 4.10: 本研究の定義で選択された変数の組み合わせ（打刻データ）

番号	打刻データを用いた変数群
1	外国語 1 年次後期
2	人文 1 年次後期
3	理科 1 年次後期
4	専門 1 年次後期
5	前期不可獲得数
6	後期不可獲得数
7	後期失格獲得数
8	1 年次 7 月打刻回数
9	1 年次 11 月打刻回数
10	1 年次 12 月打刻回数
11	1 年次 1 月打刻回数

表 4.11: 本研究の定義で選択された変数の組み合わせ（出欠データ）

番号	出欠データを用いた変数群
1	外国語 1 年次後期
2	人文 1 年次後期
3	理科 1 年次後期
4	専門 1 年次後期
5	前期不可獲得数
6	後期不可獲得数
7	後期失格獲得数
8	1 年次 11 月出席回数
9	1 年次 12 月出席回数
10	1 年次 1 月出席回数
11	1 年次 2 月出席回数
12	1 年次通年欠席回数
13	1 年次後期欠席回数

表 4.12: 打刻データと成績データの変数群：ウォード法による離散化（クラスタ数は 3）

科目	項目	1	2	3
人文 1 年次後期	離散幅	[4.00 , 3.00]	[2.00 , 2.00]	[1.00 , 0.00]
	人数	146	98	63
数学 1 年次後期	離散幅	[4.00 , 2.40]	[2.30 , 1.00]	[0.800 , 0.00]
	人数	103	149	55
理科 1 年次前期	離散幅	[4.00 , 2.50]	[2.00 , 1.50]	[1.00 , 0.00]
	人数	117	100	90
理科 1 年次後期	離散幅	[4.00 , 2.50]	[2.00 , 1.50]	[1.00 , 0.00]
	人数	95	122	90
専門 1 年次後期	離散幅	[4.00 , 2.50]	[2.40 , 2.00]	[1.75 , 0.00]
	人数	143	79	85
前期不可獲得数	離散幅	[6.00 , 2.00]	[1.00 , 1.00]	[0.00 , 0.00]
	人数	40	71	196
後期秀獲得数	離散幅	[10.0 , 4.00]	[3.00 , 2.00]	[1.00 , 0.00]
	人数	94	103	110
後期不可獲得数	離散幅	[8.00 , 3.00]	[2.00 , 1.00]	[0.00 , 0.00]
	人数	59	77	171
後期失格獲得数	離散幅	[6.00 , 2.00]	[1.00 , 1.00]	[0.00 , 0.00]
	人数	7	22	278
1 年次 5 月打刻回数	離散幅	[109 , 88]	[87 , 79]	[78 , 50]
	人数	46	95	166
1 年次 6 月打刻回数	離散幅	[108 , 82]	[81 , 70]	[69 , 40]
	人数	120	115	72
1 年次 7 月打刻回数	離散幅	[102 , 74]	[73 , 63]	[62 , 21]
	人数	151	82	74
1 年次 12 月打刻回数	離散幅	[103 , 76]	[75 , 63]	[62 , 20]
	人数	45	145	117
1 年次 1 月打刻回数	離散幅	[129 , 76]	[75 , 59]	[58 , 22]
	人数	112	154	41

表 4.13: 打刻データと成績データの変数群：ウォード法による離散化（クラスタ数は 4）

科目	項目	1	2	3	4
人文 1 年次後期	離散幅	[4.00 , 4.00]	[3.00 , 3.00]	[2.00 , 2.00]	[1.00 , 0.00]
	人数	55	91	98	63
数学 1 年次後期	離散幅	[4.00 , 2.40]	[2.30 , 1.80]	[1.70 , 1.00]	[0.800 , 0.00]
	人数	103	77	72	55
理科 1 年次前期	離散幅	[4.00 , 2.50]	[2.00 , 2.00]	[1.50 , 1.50]	[1.00 , 0.00]
	人数	117	43	57	90
理科 1 年次後期	離散幅	[4.00 , 2.50]	[2.00 , 2.00]	[1.50 , 1.50]	[1.00 , 0.00]
	人数	95	62	60	90
専門 1 年次後期	離散幅	[4.00 , 3.25]	[3.00 , 2.50]	[2.40 , 2.00]	[1.75 , 0.00]
	人数	52	91	79	85
前期不可獲得数	離散幅	[6.00 , 3.00]	[2.00 , 2.00]	[1.00 , 1.00]	[0.00 , 0.00]
	人数	19	21	71	196
後期秀獲得数	離散幅	[10.0 , 7.00]	[6.00 , 4.00]	[3.00 , 2.00]	[1.00 , 0.00]
	人数	27	67	103	110
後期不可獲得数	離散幅	[8.00 , 3.00]	[2.00 , 2.00]	[1.00 , 1.00]	[0.00 , 0.00]
	人数	59	34	43	171
後期失格獲得数	離散幅	[6.00 , 3.00]	[2.00 , 2.00]	[1.00 , 1.00]	[0.00 , 0.00]
	人数	3	4	22	278
1 年次 5 月打刻回数	離散幅	[109 , 88]	[87 , 79]	[78 , 70]	[69 , 50]
	人数	46	95	119	47
1 年次 6 月打刻回数	離散幅	[108 , 82]	[81 , 70]	[69 , 61]	[60 , 40]
	人数	120	115	45	27
1 年次 7 月打刻数	離散幅	[102 , 80]	[79 , 74]	[73 , 63]	[62 , 21]
	人数	66	85	82	74
1 年次 12 月打刻数	離散幅	[103 , 76]	[75 , 63]	[62 , 43]	[42 , 20]
	人数	45	145	106	11
1 年次 1 月打刻数	離散幅	[129 , 89]	[88 , 76]	[75 , 59]	[58 , 22]
	人数	44	68	154	41

表 4.14: 出欠データと成績データの変数群：ウォード法による離散化（クラスタ数は 3）

科目	項目	1	2	3
人文 1 年次後期	離散幅	[4.00 , 3.00]	[2.00 , 2.00]	[1.00 , 0.00]
	人数	146	98	63
数学 1 年次後期	離散幅	[4.00 , 2.40]	[2.30 , 1.00]	[0.800 , 0.00]
	人数	103	149	55
理科 1 年次後期	離散幅	[4.00 , 2.50]	[2.00 , 1.50]	[1.00 , 0.00]
	人数	95	122	90
専門 1 年次後期	離散幅	[4.00 , 2.50]	[2.40 , 2.00]	[1.75 , 0.00]
	人数	143	79	85
前期不可獲得数	離散幅	[6.00 , 2.00]	[1.00 , 1.00]	[0.00 , 0.00]
	人数	40	71	196
後期秀獲得数	離散幅	[10.0 , 4.00]	[3.00 , 2.00]	[1.00 , 0.00]
	人数	94	103	110
後期不可獲得数	離散幅	[8.00 , 3.00]	[2.00 , 1.00]	[0.00 , 0.00]
	人数	59	77	171
後期失格獲得数	離散幅	[6.00 , 2.00]	[1.00 , 1.00]	[0.00 , 0.00]
	人数	7	22	278
1 年次 7 月出席回数	離散幅	[54 , 43]	[42 , 39]	[38 , 13]
	人数	116	70	121
1 年次 12 月出席回数	離散幅	[43 , 34]	[33 , 28]	[27 , 13]
	人数	134	125	48
1 年次 1 月出席回数	離散幅	[46 , 38]	[37 , 33]	[32 , 12]
	人数	93	133	81
1 年次 2 月出席回数	離散幅	[14 , 9]	[8 , 7]	[6 , 0.00]
	人数	114	97	96
1 年次前期欠席回数	離散幅	[68 , 29]	[28 , 11]	[10 , 0]
	人数	21	77	209

表 4.15: 出欠データと成績データの変数群：ウォード法による離散化（クラスタ数は 4）

科目	項目	1	2	3	4
人文 1 年次後期	離散幅	[4.00 , 4.00]	[3.00 , 3.00]	[2.00 , 2.00]	[1.00 , 0.00]
	人数	55	91	98	63
数学 1 年次後期	離散幅	[4.00 , 2.40]	[2.30 , 1.80]	[1.70 , 1.00]	[0.800 , 0.00]
	人数	103	77	72	55
理科 1 年次後期	離散幅	[4.00 , 2.50]	[2.00 , 2.00]	[1.50 , 1.50]	[1.00 , 0.00]
	人数	95	62	60	90
専門 1 年次後期	離散幅	[4.00 , 3.25]	[3.00 , 2.50]	[2.40 , 2.00]	[1.75 , 0.00]
	人数	52	91	79	85
前期不可獲得数	離散幅	[6.00 , 3.00]	[2.00 , 2.00]	[1.00 , 1.00]	[0.00 , 0.00]
	人数	19	21	71	196
後期秀獲得数	離散幅	[10.0 , 7.00]	[6.00 , 4.00]	[3.00 , 2.00]	[1.00 , 0.00]
	人数	27	67	103	110
後期不可獲得数	離散幅	[8.00 , 3.00]	[2.00 , 2.00]	[1.00 , 1.00]	[0.00 , 0.00]
	人数	59	34	43	171
後期失格獲得数	離散幅	[6.00 , 3.00]	[2.00 , 2.00]	[1.00 , 1.00]	[0.00 , 0.00]
	人数	3	4	22	278
1 年次 7 月出席回数	離散幅	[54 , 48]	[47 , 43]	[42 , 39]	[38 , 13]
	人数	53	63	70	121
1 年次 12 月出席回数	離散幅	[43 , 34]	[33 , 32]	[31 , 28]	[27 , 13]
	人数	134	61	64	48
1 年次 1 月出席回数	離散幅	[46 , 38]	[37 , 33]	[32 , 27]	[26 , 12]
	人数	93	133	60	21
1 年次 2 月出席回数	離散幅	[14 , 11]	[10 , 9]	[8 , 7]	[6 , 0.00]
	人数	50	64	97	96
1 年次前期欠席回数	離散幅	[68 , 29]	[28 , 11]	[10 , 5]	[4 , 0]
	人数	21	77	93	116

表 4.16: 打刻データと成績データの変数群：ウォード法による離散化（クラスタ数は 3）

科目	項目	1	2	3
外国語 1 年次後期	離散幅	[4.00 , 3.50]	[3.0 , 2.40]	[2.5 , 0.50]
	人数	68	72	162
人文 1 年次後期	離散幅	[4.00 , 3.00]	[2.00 , 1.50]	[1.00 , 0.00]
	人数	145	97	60
理科 1 年次後期	離散幅	[4.00 , 2.5]	[2.00 , 1.50]	[1.00 , 0.00]
	人数	93	120	89
専門 1 年次後期	離散幅	[4.00 , 3.25]	[3.00 , 2.50]	[2.40 , 0.50]
	人数	52	89	161
前期不可獲得数	離散幅	[6.00 , 2.00]	[1.00 , 1.00]	[0.00 , 0.00]
	人数	38	70	194
後期不可獲得数	離散幅	[8.00 , 3.00]	[2.00 , 1.00]	[0.00 , 0.00]
	人数	56	77	169
後期失格獲得数	離散幅	[6.00 , 2.00]	[1.00 , 1.00]	[0.00 , 0.00]
	人数	6	21	275
1 年次 7 月打刻回数	離散幅	[102 , 74]	[73 , 63]	[62 , 21]
	人数	149	81	72
1 年次 11 月打刻回数	離散幅	[122 , 85]	[84 , 75]	[74 , 38]
	人数	122	104	76
1 年次 12 月打刻回数	離散幅	[103 , 73]	[72 , 60]	[59 , 20]
	人数	75	137	90
1 年次 1 月打刻回数	離散幅	[129 , 80]	[79 , 59]	[58 , 22]
	人数	84	178	40

表 4.17: 打刻データと成績データの変数群：ウォード法による離散化（クラスタ数は 4）

科目	項目	1	2	3	4
外国語 1 年次後期	離散幅	[4.00 , 3.50]	[3.00 , 2.90]	[2.50 , 2.40]	[2.00 , 0.50]
	人数	68	72	59	103
人文 1 年次後期	離散幅	[4.00 , 4.00]	[3.00 , 3.00]	[2.00 , 2.00]	[1.00 , 1.00]
	人数	55	90	97	60
理科 1 年次後期	離散幅	[4.00 , 2.50]	[2.00 , 1.40]	[1.50 , 1.40]	[1.00 , 0.00]
	人数	93	62	58	89
専門 1 年次後期	離散幅	[4.00 , 3.25]	[3.00 , 2.50]	[2.4 , 2.00]	[1.75 , 0.50]
	人数	52	89	78	83
前期不可獲得数	離散幅	[6.00 , 3.00]	[2.00 , 2.00]	[1.00 , 1.00]	[0.00 , 0.00]
	人数	17	21	70	194
後期不可獲得数	離散幅	[8.00 , 3.00]	[2.00 , 2.00]	[1.00 , 1.00]	[0.00 , 0.00]
	人数	56	34	43	169
後期失格獲得数	離散幅	[6.00 , 3.00]	[2.00 , 2.00]	[1.00 , 1.00]	[0.00 , 0.00]
	人数	3	3	21	275
1 年次 7 月打刻数	離散幅	[102 , 80]	[79 , 74]	[73 , 63]	[62 , 21]
	人数	65	84	81	72
1 年次 11 月打刻数	離散幅	[122 , 85]	[84 , 75]	[74 , 66]	[65 , 38]
	人数	122	104	47	29
1 年次 12 月打刻数	離散幅	[103 , 73]	[72 , 60]	[59 , 43]	[42 , 20]
	人数	75	137	80	10
1 年次 1 月打刻数	離散幅	[129 , 80]	[79 , 72]	[71 , 59]	[58 , 22]
	人数	84	66	112	40

表 4.18: 出欠データと成績データの変数群：ウォード法による離散化（クラスタ数は 3）

科目	項目	1	2	3
外国語 1 年次後期	離散幅	[4.00 , 3.50]	[3.0 , 2.40]	[2.5 , 0.50]
	人数	68	72	162
人文 1 年次後期	離散幅	[4.00 , 3.00]	[2.00 , 1.50]	[1.00 , 0.00]
	人数	145	97	60
理科 1 年次後期	離散幅	[4.00 , 2.5]	[2.00 , 1.50]	[1.00 , 0.00]
	人数	93	120	89
専門 1 年次後期	離散幅	[4.00 , 3.25]	[3.00 , 2.50]	[2.40 , 0.50]
	人数	52	89	161
前期不可獲得数	離散幅	[6.00 , 2.00]	[1.00 , 1.00]	[0.00 , 0.00]
	人数	38	70	194
後期不可獲得数	離散幅	[8.00 , 3.00]	[2.00 , 1.00]	[0.00 , 0.00]
	人数	56	77	169
後期失格獲得数	離散幅	[6.00 , 2.00]	[1.00 , 1.00]	[0.00 , 0.00]
	人数	6	21	275
1 年次 11 月出席回数	離散幅	[56 , 46]	[45 , 40]	[39 , 20]
	人数	149	104	49
1 年次 12 月出席回数	離散幅	[43 , 34]	[33 , 28]	[27 , 13]
	人数	132	123	47
1 年次 1 月出席回数	離散幅	[46 , 38]	[37 , 33]	[32 , 12]
	人数	93	130	79
1 年次 2 月出席回数	離散幅	[14 , 10]	[9 , 7]	[6 , 0.00]
	人数	84	124	94
1 年次通年欠席回数	離散幅	[130 , 34]	[33 , 18]	[17 , 1]
	人数	93	78	131
1 年次後期欠席回数	離散幅	[81 , 26]	[25 , 7]	[6 , 0.00]
	人数	76	151	75

表 4.19: 出欠データと成績データの変数群：ウォード法による離散化（クラスタ数は 4）

科目	項目	1	2	3	4
外国語 1 年次後期	離散幅	[4.00 , 3.50]	[3.00 , 2.90]	[2.50 , 2.40]	[2.00 , 0.50]
	人数	68	72	59	103
人文 1 年次後期	離散幅	[4.00 , 4.00]	[3.00 , 3.00]	[2.00 , 2.00]	[1.00 , 1.00]
	人数	55	90	97	60
理科 1 年次後期	離散幅	[4.00 , 2.50]	[2.00 , 1.40]	[1.50 , 1.40]	[1.00 , 0.00]
	人数	93	62	58	89
専門 1 年次後期	離散幅	[4.00 , 3.25]	[3.00 , 2.50]	[2.40 , 2.00]	[1.75 , 0.50]
	人数	52	89	78	83
前期不可獲得数	離散幅	[6.00 , 3.00]	[2.00 , 2.00]	[1.00 , 1.00]	[0.00 , 0.00]
	人数	17	21	70	194
後期不可獲得数	離散幅	[8.00 , 3.00]	[2.00 , 2.00]	[1.00 , 1.00]	[0.00 , 0.00]
	人数	56	34	43	169
後期失格獲得数	離散幅	[6.00 , 3.00]	[2.00 , 2.00]	[1.00 , 1.00]	[0.00 , 0.00]
	人数	3	3	21	275
1 年次 11 月出席回数	離散幅	[56 , 49]	[48 , 46]	[45 , 40]	[39 , 20]
	人数	88	61	104	49
1 年次 12 月出席回数	離散幅	[43 , 34]	[33 , 32]	[31 , 28]	[27 , 13]
	人数	132	59	64	47
1 年次 1 月出席回数	離散幅	[46 , 42]	[41 , 38]	[37 , 33]	[32 , 12]
	人数	28	65	130	79
1 年次 2 月出席回数	離散幅	[14 , 10]	[9 , 8]	[7 , 7]	[6 , 0.00]
	人数	84	70	54	94
1 年次通年欠席回数	離散幅	[130 , 57]	[56 , 34]	[33 , 18]	[17 , 1]
	人数	31	62	78	131
1 年次後期欠席回数	離散幅	[81 , 51]	[50 , 26]	[25 , 7]	[6 , 0.00]
	人数	15	61	151	75

4.1.5　発見の評価方法

本研究では教員にかかる指導のコストを削減することを目的としている. そのため指導の対象者をできるだけ減らしたい. しかし, 単純に対象者を減らすだけでは, 肝心の要注意学生を多数見逃してしまい, 本末転倒となってしまう. 逆に要注意学生を全員見つけようとして, 指導対象者をいたずらに増やすと, 今度は教員にかかる指導のコストが大幅に増えてしまう. 指導のコストと要注意学生の発見のバランスを考慮した評価をする必要がある.

本研究では機械学習の評価方法を用いていて, 要注意学生であるかどうかの発見を想定し, 評価方法を説明する. 実際の学生の中には要注意学生である学生とそうでない学生が存在し, そして各学生に対し要注意学生であるかどうかの予測を与える. ここで実際に要注意学生であることを Positive な事象（本来の意味での Positive ではなく, あくまで本研究においてである）と考えたとき,「要注意学生である」という予測を与えられた場合,Pisitive な事象が True（正解）であるということから True Positive（以下 TP）と表す. このように他の事象にも当てはめると, 実際の要注意学生に対して「要注意学生ではない」という予測が与えられた場合は False Positive（以下 FP）, 実際は要注意学生ではない学生に対して「要注意学生である」という予測が与えられた場合は False Negative（以下 FN）, 実際は要注意学生ではない学生に対して「要注意学生ではない」という予測が与えられた場合は True Negative（以下 TN）とする. これらをまとめたものを表 4.20 に示す. これらを用いて, 正解率と再現率, 適合率, そして発見精度の評価指標である F-measure を求める. これらの指標は以下の式 4.1〜4.4 のように求められる.

表 4.20: 実際と予測結果に応じた表記

	実際に『要注意学生』である	実際に『要注意学生』でない
「『要注意学生』である」と予測	True Positive（TP）	False Negative（FN）
「『要注意学生』でない」と予測	False Positive（FP）	True Negative（TN）

$$\text{正解率} = \frac{TP+TN}{TP+FP+FN+TN} \tag{4.1}$$

$$\text{再現率} = \frac{TP}{TP+FP} \tag{4.2}$$

$$\text{適合率} = \frac{TP}{TP+FN} \tag{4.3}$$

$$F-measure = \frac{2\,\text{再現率} * \text{適合率}}{\text{再現率}+\text{適合率}} \tag{4.4}$$

正解率は全体の的中率を表している. 再現率とは実際の要注意学生のうち, 何人の要注意学生を予測できたかを表している. 適合率とは要注意学生と予測した学生のうち, 何人が実際

の要注意学生であったかを表している. F-measure は適合率と再現率の調和平均である.

例として今 20 名の学生がいるとして, この 20 名の中に要注意学生が 5 名いたとする. 予測によって 8 名の学生を要注意学生と予測し, その予測された 8 名のうち 3 名が要注意学生であったとする. この場合において上で述べた 4 つの値は, 次のように求められる.

$$\text{正解率} = \frac{3+10}{20} = 0.65 \tag{4.5}$$

$$\text{再現率} = \frac{3}{5} = 0.6 \tag{4.6}$$

$$\text{適合率} = \frac{3}{8} = 0.375 \tag{4.7}$$

$$F - measure = \frac{2 * 0.375 * 0.6}{0.375 + 0.6} = 0.46 \tag{4.8}$$

この結果は「全体としての正解率は 65 %である」「実際の要注意学生の 60 %を拾い上げた」「要注意学生と予測された学生のうち実際に要注意学生であったのは 37.5 %で, そうでない学生が 62.5 %である」と評価できる. 本研究の欠点として, 実際に要注意学生であるにもかかわらず「要注意学生ではない」という予測（FP）が出てしまったり, 実際には要注意学生ではないが「要注意学生である」という予測（FN）がでてしまう可能性が大なり小なりあることである. そこで本研究では前者の FP をできるだけ少なくしたいと考えているので, 再現率の評価に重きを置いた.

実際に例を挙げて予測と再現率に重きを置いた評価を考える. 図 4.1 にその例を示す. 全学生が 12 名であり, △マークが要注意学生, ○マークが要注意学生ではない学生とする. 赤の枠で囲まれた学生が要注意学生であると予測された学生である. 左の例では 3 名が要注意学生であると予測され, 実際には 1 名が要注意学生であった. 右の例では 8 名が要注意学生であると予測され, 4 名が実際に要注意学生であった. この時, どちらの例が本研究において優れた予測かを説明する. 左の例において正解率は 75 %, 再現率は 50 %, 適合率は 67 %となっている. また, 右の例において正解率は 67 %, 再現率は 100 %, 適合率は 50 %となっている. 正解率と適合率に注目すると, 左の例のほうが両方とも値が高いので優れていると考えられる. しかし, 本研究では FP をできるだけ少なくしたいと考えており, 再現率に重きを置いている. では再現率をみてみると左の例が 50 %であるのに対して, 右の例では 100 %であり, 再現率に重きを置いた場合は右の例のほうが優れた予測と言える. しかし, 再現率にとらわれすぎてしまうのはいけない. 先にも述べたように本研究では教員の指導にかかるコストを削減することも目的の一つである. 極端な話であるが, 全学生を要注意学生であるとしてしまえば, 再現率は 100 %にすることができる. これでは教員にかかる指導のコストを削減することができない. この場合では F-meansure が役に立つ. 図 4.1 を用いて左の例と全学生を要注意学生であると予測した場合を比較する. 全学生を要注意学生であると予測した場合, 正解率は 33 %, 再現率は 100 %, 適合率は 33 %となる. 先ほどと同じように再現率に注目して評価すると, 左の例では 50 %に対して 100 %であり, 全学生を要注意学生と予測した場合のほうが優れた予測となってしまう. しかし F-measure を求めると 0.496 となる. 左の例は 0.573 である. ゆえに予測の評価をする際には F-meansure の値も考慮する必要がある.

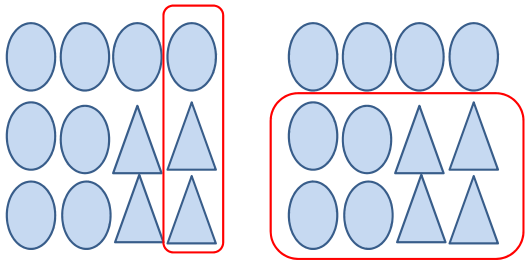

図 4.1: 予測の例

4.1.6　発見モデルの評価

本研究ではベイジアンネットワークの確率推論により, 要注意学生である確率を求めることで予測する. ベイジアンネットワークの特徴として, ある事象の事後確率を算出している点である. 要注意学生であるかどうかの予測において,「要注意学生である確率が 80 %」であれば,「要注意学生ではない確率は 20 %」といった形で表される. なので「確率が一定の値以上になれば要注意学生とする」というように閾値を設定することで, 柔軟な予測を行うことができる. 例えば上記の例で言えば,「要注意学生である確率が 80 %」であるので, 閾値を 70 %に設定すれば要注意学生であると予測される. しかしこの学生の他に「要注意学生である確率が 50 %」学生がいたとすると, 閾値が 70 %では要注意学生であると予測してくれない. この場合閾値を 40 %に設定することで, この学生も要注意学生であると予測することができる. 本研究では閾値を 50 %,30 %, 事前確率に設定し, その発見精度を検証した. 事前確率は全体の学生に対する要注意学生の割合で求めることができる. 本研究での予測の対象者は 302 名で, 要注意学生は 25 名であるため, 事前確率は $25 \div 302 = 0.083$ なので 8.3 %となる. また, 従来の予測対象者は 307 名で, 要注意学生は 41 名であるため, 事前確率は $41 \div 307 = 0.134$ なので 13.4 %となる. 事後確率が事前確率より上がったときに要注意学生と判定するため, 事前確率も判定の閾値とする.　また各モデルにおける制度の評価方法は,leave one out 法も用いた.

4.2　要注意学生の発見

ここまで要注意学生の厳密な定義や, 用いる変数の組み合わせと取捨選択, その変数の離散化, 評価方法などの予測のための準備をした. 本節では, ここまで説明してきたことを用いて要注意学生の予測を行う. 本研究では, 出欠データと成績データを用いた要注意学生の発見モデルを提案している. この発見モデルの有用性を示すため, 打刻データによる要注意学生の予測と比較する. また, 従来の要注意学生の定義と本研究での要注意学生の定義の発見モデルの比較も行う.

4.2.1　従来の定義の要注意学生の発見

打刻データと成績データの属性変数を 36 変数として,CFS による属性選択を行い上記した表 4.8 の変数が選択された．その変数をウォード法によって表 4.12 と表 4.13 のように離散化した．出欠データと成績データの属性変数を 39 変数として,CFS による属性選択を行い上記した表 4.9 の変数が選択された．その変数をウォード法によって表 4.14 と表 4.15 のように離散化した．この変数をそれぞれ用いて, ベイジアンネットワークによる発見モデルを構築した．

打刻データと成績データの変数をウォード法によってクラスタ数を 3 で離散化された発見モデル

ウォード法によって 14 個の変数を離散化した．その時各変数をクラスタ数が 3 とした場合の発見モデルについて説明する．有効グラフ構造については,Naive Bayes 構造によるモデルを構築した．モデルの構造を図 4.2 に示す．そして閾値を 50 %,30 %, 事前確率である 13.4 %にした場合の発見精度を表 4.21 に示す．

正解率は閾値を 50 %に設定した時が最もよく, 再現率においては閾値を 13.4 %に設定した時が最もよい結果となった．しかし F-measure の値を見てみると, 閾値を 30 %に設定した時が一番高かった．これは他の閾値の場合と比べて, 要注意学生の発見と, 指導にかかるコストのバランスが取れていると言える．

表 4.21: クラスタ数 3 で離散化したモデルの精度一覧（打刻データ）

閾値	正解率			再現率			適合率			F-measure
	対象	的中		対象	的中		対象	的中		
50 %	307	247	80 %	41	20	49 %	59	20	34 %	0.400
30 %	307	240	78 %	41	24	59 %	74	24	32 %	0.417
13.4. %	307	225	73 %	41	26	63 %	93	26	26 %	0.388

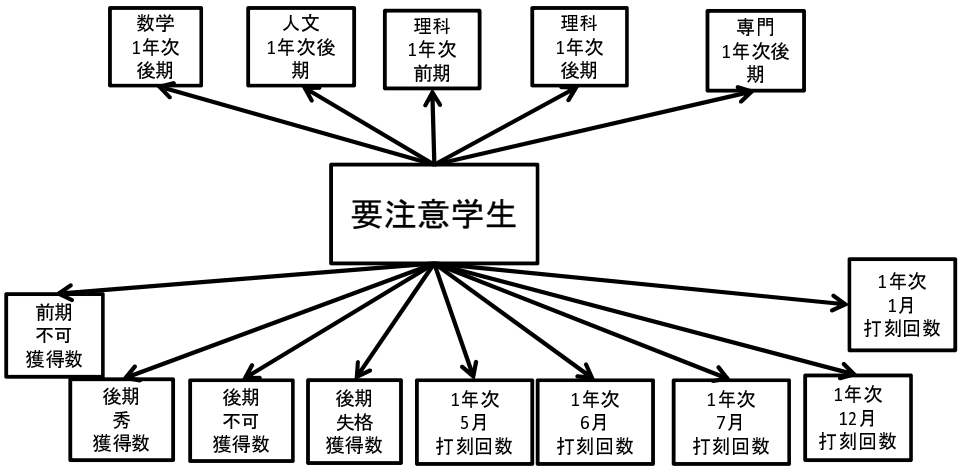

図 4.2: 打刻データと成績データの変数で構築されたモデル

打刻データと成績データの変数をウォード法によってクラスタ数を 4 で離散化された発見モデル

ウォード法によって 14 個の変数を離散化した. その時各変数をクラスタ数が 4 とした場合の発見モデルについて説明する. 有効グラフ構造については,Naive Bayes 構造によるモデルを構築した. モデルの構造は図 4.2 と同じため省略する. そして閾値を 50 %,30 %, 事前確率である 13.4 %にした場合の発見精度を表 4.22 に示す.

F-measure の値はすべて 0.4 を超えていて, 最も高いのは閾値が 30 %のときであった. しかしながら閾値が 13.4 の時の再現率は 68 %と 28 名の要注意学生を発見することができていた. また, 閾値 30 %から 13.4 にかえて増えた指導対象者は 11 名, その中に要注意学生は 2 名である. 閾値が 13.4 %のモデルは, 対象 307 名中 91 名を指導対象とすることで, GPA1.0 以上の学生から 68 %の要注意学生を発見することができている.

表 4.22: クラスタ数 4 で離散化したモデルの精度一覧（打刻データ）

	正解率			再現率			適合率			F-measure
閾値	対象	的中		対象	的中		対象	的中		
50 %	307	248	81 %	41	20	49 %	58	20	34 %	0.404
30 %	307	238	78 %	41	26	63 %	80	26	33 %	0.430
13.4 %	307	231	75 %	41	28	68 %	91	28	31 %	0.424

出欠データと成績データの変数をウォード法によってクラスタ数を 3 で離散化された発見モデル

ウォード法によって 14 個の変数を離散化した. その時各変数をクラスタ数が 3 とした場合の発見モデルについて説明する. 有効グラフ構造については,Naive Bayes 構造によるモデルを構築した. モデルの構造を図 4.3 に示す. そして閾値を 50 %,30 %, 事前確率である 13.4 %にした場合の発見精度を表 4.23 に示す.

再現率は閾値を 50 %に設定した時が最もよく, 適合率は閾値を 13.4 %に設定した時に最も良い値であった. F-measure の値はどの閾値でも大きな差はなかった. これは, 発見した要注意学生の人数と, 指導にかかるコストが, どの閾値でも同じぐらいのバランスであったと考えられる.

表 4.23: クラスタ数 3 で離散化したモデルの精度一覧（出欠データ）

閾値	正解率			再現率			適合率			F-measure
	対象	的中		対象	的中		対象	的中		
50 %	307	249	81 %	41	19	46 %	55	19	35 %	0.395
30 %	307	240	78 %	41	21	51 %	68	21	31 %	0.385
13.4 %	307	227	74 %	41	25	61 %	89	25	28 %	0.384

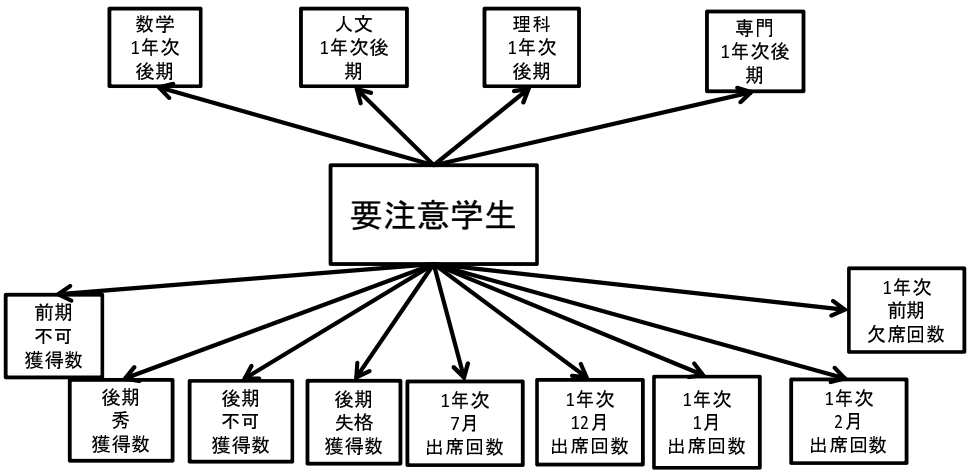

図 4.3: 打刻データと成績データの変数で構築されたモデル

出欠データと成績データの変数をウォード法によってクラスタ数を 4 で離散化された発見モデル

ウォード法によって 14 個の変数を離散化した．その時各変数をクラスタ数が 4 とした場合の発見モデルについて説明する．有効グラフ構造については,Naive Bayes 構造によるモデルを構築した．モデルの構造は図 4.3 と同じため省略する．そして閾値を 50 %,30 %, 事前確率である 13.4 %にした場合の発見精度を表 4.24 に示す.

この結果も出欠データと成績データの変数をウォード法によってクラスタ数を 3 で離散化された発見モデルと同じ傾向が見受けられる. しかし, 閾値 13.4 %の場合に指導対象者が 86 名と 3 名減っている．そのためわずかであるが F-measure の値が向上しているのが見て取れる．

表 4.24: クラスタ数 4 で離散化したモデルの精度一覧（出欠データ）

	正解率			再現率			適合率			F-measure
閾値	対象	的中		対象	的中		対象	的中		
50 %	307	249	81 %	41	19	46 %	55	19	35 %	0.395
30 %	307	240	78 %	41	21	51 %	68	21	31 %	0.385
13.4 %	307	230	75 %	41	25	61 %	86	25	29 %	0.394

4.2.2　本研究で定義した要注意学生の発見

打刻データと成績データの属性変数を 36 変数として,CFS による属性選択を行い上記した表 4.8 の変数が選択された．その変数をウォード法によって表 4.16 と表 4.17 のように離散化した．出欠データと成績データの属性変数を 39 変数として,CFS による属性選択を行い上記した表 4.9 の変数が選択された．その変数をウォード法によって表 4.18 と表 4.19 のように離散化した．この変数をそれぞれ用いて, ベイジアンネットワークによる発見モデルを構築した.

打刻データと成績データの変数をウォード法によってクラスタ数を 3 で離散化された発見モデル

ウォード法によって 11 個の変数を離散化した．その時各変数をクラスタ数が 3 とした場合の発見モデルについて説明する．有効グラフ構造については,Naive Bayes 構造によるモデルを構築した．モデルの構造を図 4.4 に示す．そして閾値を 50 %,30 %, 事前確率である 8.3 %にした場合の発見精度を表 4.25 に示す.

正解率がもっとも高い値を示したのは閾値が 50 %の場合であった．再現率では閾値が 8.3 %の時に一番よい値であった. 閾値が 50 %と 30 %を比較してみると, 再現率が変わらないのに, 要注意学生であると予測した人数が閾値が 30 %のほうが多いため, F-measure の値は

閾値が 50 %のほうが高くなっている. この場合は閾値を 50 %にした場合がより優れた予測と言える.

表 4.25: クラスタ数 3 で離散化したモデルの精度一覧（打刻データ）

閾値	正解率			再現率			適合率			F-measure
	対象	的中		対象	的中		対象	的中		
50 %	302	260	86 %	25	14	56 %	45	14	31 %	0.400
30 %	302	256	85 %	25	14	56 %	51	14	27 %	0.368
8.3 %	302	233	77 %	25	16	64 %	76	16	21 %	0.317

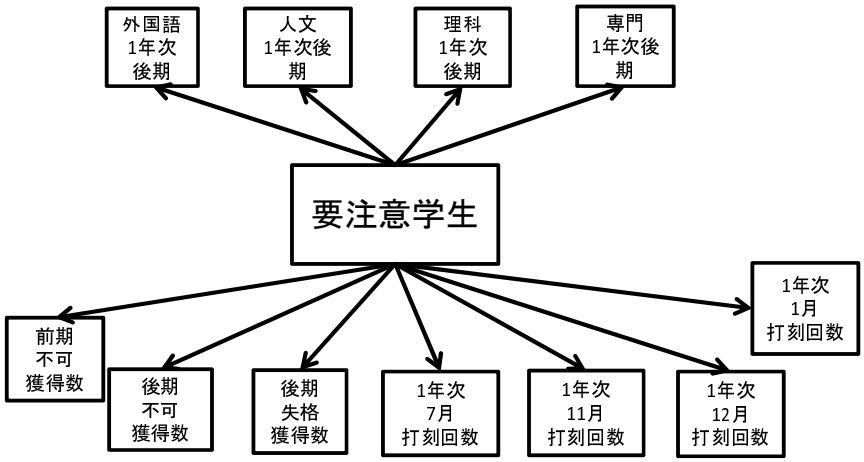

図 4.4: 打刻データと成績データの変数で構築されたモデル

打刻データと成績データの変数をウォード法によってクラスタ数を 4 で離散化された発見モデル

ウォード法によって 11 個の変数を離散化する時に, クラスタ数を 4 にした場合の発見モデルについて述べる. この発見モデルについても有効グラフ構造は Native Bayes 構造によるモデルを構築した. モデルの構造は図 4.4 と同じため省略する. そして閾値を 50 %,30 %, 事前確率である 8.3 %にした場合の発見精度を表 4.26 に示す.

F-measure の値が最も高かったのは閾値が 50 %のときであった. しかし再現率を見てみると閾値が 50 %の時 52 %であるのに対して, 閾値が 8.3 %では 64 %と 12 %も違い, 人数で

は 3 人の違いが確認できる. 閾値が 8.3 %のモデルは, 対象 302 名中 72 名を指導対象とすることで, GPA1.0 以上の学生から 64 %の要注意学生を発見することができている.

表 4.26: クラスタ数 4 で離散化したモデルの精度一覧（打刻データ）

閾値	正解率			再現率			適合率			F-measure
	対象	的中		対象	的中		対象	的中		
50 %	302	264	87 %	25	13	52 %	39	13	33 %	0.406
30 %	302	258	85 %	25	13	52 %	46	13	28 %	0.366
8.3 %	302	237	78 %	25	16	64 %	72	16	22 %	0.330

出欠データと成績データの変数をウォード法によってクラスタ数を 3 で離散化された発見モデル

ウォード法によって 13 個の変数を離散化した. その時各変数をクラスタ数が 3 とした場合の予測モデルについて説明する. 有効グラフ構造については,Naive Bayes 構造によるモデルを構築した. モデルの構造を図 4.5 に示す. そして閾値を 50 %,30 %, 事前確率である 8.3 %にした場合の発見精度を表 4.27 に示す.

F-measure の値は閾値を 50 %に設定したモデルが 0.418 と最もよかった. 再現率に注目すると, 閾値を 8.3 %に設定したモデルが 25 名中 18 名で 72 %も予測することができている. しかし指導の対象者が閾値が 50 %の場合では 42 名に対して, 閾値が 8.3 %の場合では 72 名となっており, 30 名増えていることがわかる.

表 4.27: クラスタ数 3 で離散化したモデルの精度一覧（出欠データ）

閾値	正解率			再現率			適合率			F-measure
	対象	的中		対象	的中		対象	的中		
50 %	302	263	87 %	25	14	56 %	42	14	33 %	0.418
30 %	302	256	85 %	25	15	60 %	51	15	29 %	0.395
8.3 %	302	241	81 %	25	18	72 %	72	16	25 %	0.371

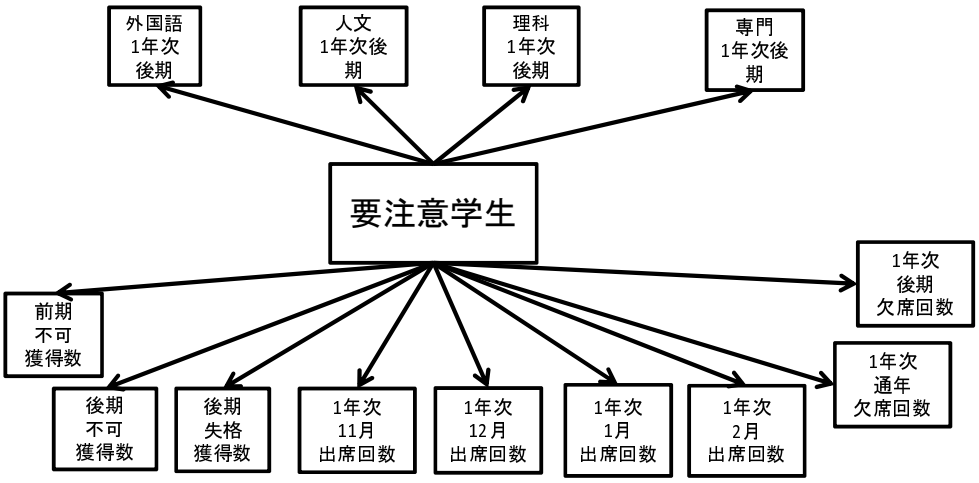

図 4.5: 出欠データと成績データの変数で構築されたモデル

出欠データと成績データの変数をウォード法によってクラスタ数を 4 で離散化された発見モデル

ウォード法によって 13 個の変数を離散化する時に, クラスタ数を 4 にした場合の予測モデルについて述べる. この予測モデルについても有効グラフ構造は Native Bayes 構造によるモデルを構築した. モデルの構造は図 4.5 と同じため省略する. そして閾値を 50 %,30 %, 事前確率である 8.3 %にした場合の発見精度を表 4.28 に示す.

正解率が最もよかったのは閾値を 50 %に設定したモデルであった. 適合率では閾値が 50 %と 30 %の時が最もよく, 指導対象者を絞れている. 再現率において閾値が 30 %の場合と 8.3 %の場合が優れていて, それぞれ 68 %,72 %となっている. さらに閾値を 30 %に設定した場合の F-measure の値は 0.447 とここまでで最も良い値をとっている.

表 4.28: クラスタ数 4 で離散化したモデルの精度一覧（出欠データ）

閾値	正解率			再現率			適合率			F-measure
	対象	的中		対象	的中		対象	的中		
50 %	302	263	87 %	25	14	56 %	42	14	33 %	0.418
30 %	302	260	86 %	25	17	68 %	51	17	33 %	0.447
8.3 %	302	244	81 %	25	18	72 %	69	18	26 %	0.383

4.3　要注意学生の発見の結論

本章では発見のための下準備と要注意学生の発見を行った. まず要注意学生の厳密な定義を行い予測する対象を絞り, 要注意学生を明確にした. その後, 予測に用いる変数を組み合わせ,CFS による変数選択を行い, 変数を取捨選択した. ベイジアンネットワークでは, 変数は原則離散化されていなければならない. 本研究で用意した変数はすべて数値で表されているので, 離散化を行った. 発見の評価方法を検討し, 要注意学生をできるだけ多く発見することと, 教員が指導にかかるコストの削減の両者をバランスよくできている発見モデルを評価する. これらの下準備を経て, 要注意学生の発見を行った. 従来までの要注意学生の定義と本研究で用いた要注意学生の定義の比較検証と, 出欠データの有用性を示すために, 打刻データでの発見と比較を行う.

以下の表 4.29 に本研究で行った発見モデルを再現率の高い順に示し, 再現率が同じ値の場合は, 適合率が高い方を順位が上になるようにしている. また, 再現率が 60 %未満のモデルは省略している. 再現率が最もよかった発見モデルは, 本研究で定義した要注意学生で, 出欠データと成績データを用いたモデルであった. 25 名中 18 名を見つけることに成功している. また従来の要注意学生の定義で最もよい再現率は 68 %であり, 同じ再現率の値に, 本研究で定義した要注意学生で, 出欠データと成績データを用いたモデルがある. このモデルは従来の定義のモデルと再現率は同じであるが, 適合率が高くなっている. つまり, 本研究のモデルは, 従来の研究のモデルと同じ割合の要注意学生を, 指導コストを減らしつつ, 発見することができる. また上位には, 本研究で定義した要注意学生のモデルが多数存在していることが見て取れる. しかしながら, 従来の要注意学生の定義において, 出欠データと成績データを用いたモデルが, 打刻データと成績データを用いたモデルよりも再現率が低いという結果が出ている. さらに, 適合率においても, 必ずしも優れているとは言い難い結果となっている. 以上のことを考慮すると, 本研究で定義した要注意学生の発見モデルにおいて, 出欠データは有用性があるということができる. また, 本研究では, 要注意学生をできるだけ発見し, なおかつ教員の指導にかかるコストを削減することを目的としている. 本研究で定義した要注意学生, 出欠データと成績データ, クラスタ数 4, 閾値 8.3 %のモデルが本研究での目的に最も即したモデルであるとし, このモデルを本研究での要注意学生の発見モデルとして採用する.

この発見モデルによる具体的な修学指導をシミュレートすると, 「1 年前期・後期の GPA が 1.0 より高い学生」302 名から, 発見モデルによって予測された 69 名の学生が修学指導対象者となる. 69 名中 18 名が要注意学生として存在する. 本研究における発見の評価を表 4.30 に示す. したがって, 本研究における発見モデルの成果は以下の式に示す.

$$\text{正解率} = \frac{TP+TN}{TP+FP+FN+TN} = \frac{18+226}{302} = 80.8\ \% \tag{4.9}$$

$$\text{再現率} = \frac{TP}{TP+FP} = \frac{18}{25} = 72\ \% \tag{4.10}$$

$$\text{適合率} = \frac{TP}{TP+FN} = \frac{18}{69} = 26\ \% \tag{4.11}$$

$$F-measure = \frac{2\,\text{再現率} * \text{適合率}}{\text{再現率}+\text{適合率}} = \frac{2*0.72*0.26}{0.72+0.26} = 0.383 \tag{4.12}$$

表 4.29: 再現率の値が高かったモデル

順位	再現率	適合率	F-measure	手法
1	72 %	26 %	0.383	新・出欠データ・クラスタ数 4・閾値 8.3 %
2	72 %	25 %	0.371	新・出欠データ・クラスタ数 3・閾値 8.3 %
3	68 %	33 %	0.447	新・出欠データ・クラスタ数 4・閾値 30 %
4	68 %	31 %	0.424	旧・打刻データ・クラスタ数 4・閾値 13.4 %
5	64 %	22 %	0.330	新・打刻データ・クラスタ数 4・閾値 8.3 %
6	64 %	21 %	0.317	新・打刻データ・クラスタ数 3・閾値 8.3 %
7	63 %	33 %	0.430	旧・打刻データ・クラスタ数 4・閾値 30 %
8	63 %	26 %	0.388	旧・打刻データ・クラスタ数 3・閾値 13.4 %
9	61 %	28 %	0.384	旧・出欠データ・クラスタ数 3・閾値 13.4 %
10	61 %	28 %	0.384	旧・出欠データ・クラスタ数 3・閾値 13.4 %
11	60 %	29 %	0.395	新・出欠データ・クラスタ数 3・閾値 30 %

※従来の要注意学生の定義を「旧」, 本研究での定義を「新」と表記

表 4.30: 本研究で採用したモデルを用いた時の各人数

	実際に『要注意学生』である	実際に『要注意学生』でない	合計
「『要注意学生』である」と予測	18（TP）	51（FN）	69
「『要注意学生』でない」と予測	7（FP)	226（TN）	233
合計	25	277	302

すなわち,302 名の学生を指導する場合の約 4 分の 1 のコストで, 要注意学生の 72 %を発見し, 指導することができることを示した. さらに, 予測が与えられた各学生の実際の 1 年次通年の GPA を調査した. 図 4.6 に、GPA のヒストグラムを示す. True Positive と False Negative のヒストグラムより, 本定義における要注意学生は GPA の低い学生がほとんどであるが, True Positive において,5 名中 4 名の GPA2.0 の学生や,GPA3.0 の学生を要注意学生として発見することができている. さらに,True Negative のヒストグラムから,1 年次通年 GPA が 2.0 未満の学生の存在も確認できた. つまり,GPA が良くても今後の修学状況が悪化してしまう学生を拾い上げ, GPA が悪くても今後の修学状況に支障が無い学生を指導対象者と見なしていないことが分かった. しかしながら,False Positive において,GAP2.0 以上の学生が少数ながらも存在している. 彼らを要注意学生として発見するにはどうすればよいのか, さらには要注意学生となる理由について, 今後追究する必要がある.

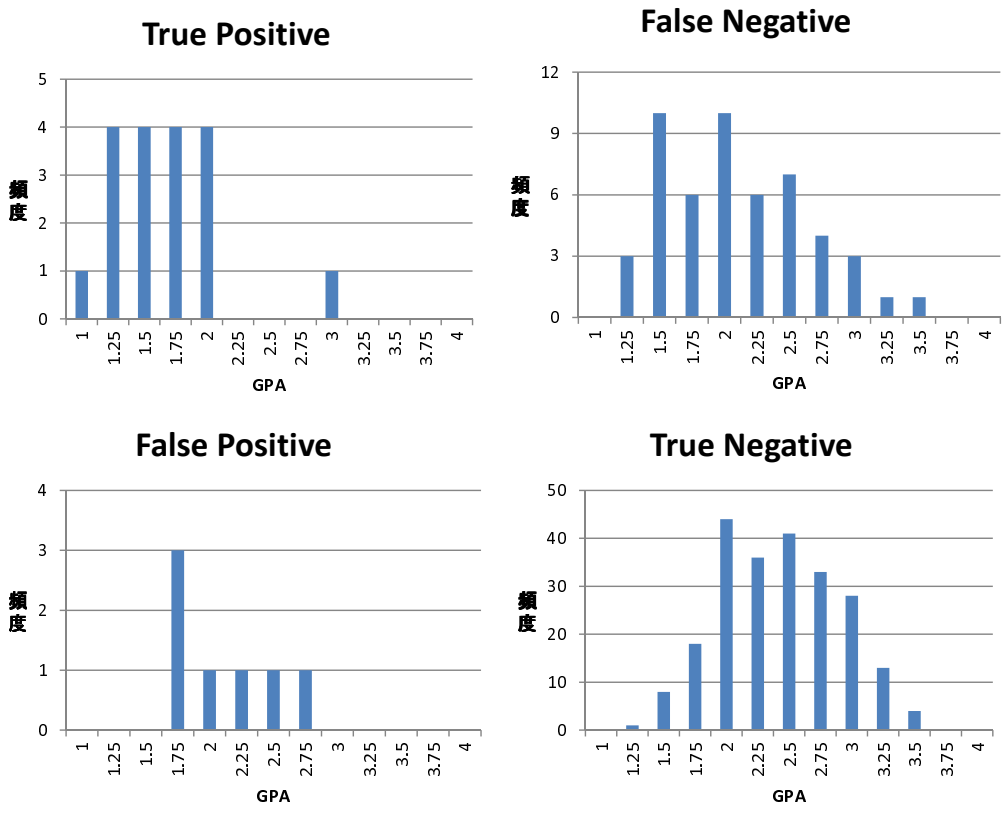

図 4.6: 各集合の GPA の分布

第5章　むすび

本研究では, 要注意学生の発見において打刻データの代わりに出欠データを使用することの有用性を示した. 第 2 章では本研究で用いた手法の理論について説明し, 第 3 章では要注意学生の定義や発見に用いたデータの概要や拡張及び変換を行った. そして第 4 章では, 要注意学生の厳密な定義や第 3 章で説明したデータの用い方, 最後に要注意学生の予測を行い, 打刻データの代わりに出欠データを使用した発見の有用性を示した.

従来の要注意学生の定義のまま「1 年次の GPA が前期か後期のどちらか一方でも 1.0 を下回っている学生」としてしまうと, 就職活動や大学院入試などによる計画的な留年をした学生や, 転学などの積極的な理由による学生までも要注意学生としてしまう. そこで 1 年次の GPA1.0 を上回っている学生を調査し, 要注意学生のさらなる絞り込みを行った. 調査の結果要注意学生の定義は以下のようにするのが良いと判断できた. この定義のもとに発見の対象者と要注意学生の絞り込み発見を行う.

- GPA が 1.0 以上で留年または退学してしまう学生
- 3 年間で卒業研究に着手した者は要注意学生から外す
- 入学から 3 年以内に退学した者はデータから除外する

出欠データと成績データによるベイジアンネットワークによる発見モデルの有用性を示すために, 打刻データと成績データによるベイジアンネットワークによる発見モデルとの比較を行った. その結果, 再現率において出欠データを用いた発見モデルのほうが優れていることがわかった. 本研究では, 要注意学生をできるだけ予測し, かつ指導にかかるコストを削減することを目的としている. それゆえに, 再現率において優れている出欠データと成績データの発見モデルに有用性があることがわかった. また, 本研究での要注意学生の定義と, 従来までの要注意学生の定義では, 本研究での定義のほうが再現率において高い値を示すことがわかった.　しかし出欠データと成績データによるベイジアンネットワークによる発見モデルの有用性は, あくまで本研究で定義した要注意学生での有用性となっている. 今回の定義よりも絞り込んだ定義の時や, またもっと曖昧な定義での時では, 有用性が必ずしも保証される訳でない. また, 本研究で用いたデータは過去のある年度におけるデータである. 違う年度においては要注意学生の特徴が違っていたりすることが考えられ, 定義自体を変える必要があるかもしれない. そのため今後は, さらに違う年度を増やしより普遍的な要注意学生の定義やその発見を行う必要があり, 発見モデルを他の年度に適用しても十分な精度を示すモデルの構築をしなければならない. これらの課題をクリアし, 最終的には実用化できるようなモデルの構築を目指す.

謝辞

本研究を進めるにあたって，日頃から多大な御尽力を頂き，ご指導を賜りました名古屋工業大学, 舟橋健司　准教授, 伊藤宏隆　助教　に心から感謝致します.

また, 本研究の実験のためのデータの提供元である, 出欠システム及びコースマネージメントシステムの開発に尽力されました, 名古屋工業大学情報基盤センター長　松尾啓志　教授, 内匠逸　教授, 情報基盤センター教職員の皆様に心から感謝いたします.

最後に, 本研究に多大な御協力頂きました舟橋研究室諸氏に心から感謝致します.

参考文献

[1] 伊藤宏隆, 舟橋健司, 中野智文, 内匠逸, 松尾啓志, 大貫徹, “名古屋工業大学における Moodle の構築と運用”, メディア教育研究, 4 巻, 2 号, 15-21 (2008)

[2] 伊藤暁人,“ニューラルネットワークによる学生の成績予測とその学習指導への適用可能性の検討”, 平成 22 年度名古屋工業大学卒業研究論文 (2010)

[3] 伊藤雄真,“IC カード打刻データと修学データを用いた学生の将来の学習レベル予測と特徴分析”, 平成 24 年度名古屋工業大学卒業研究論文 (2012)

[4] 原圭司, 高橋健一, 上田祐彰, “ベイジアンネットワークを用いた授業アンケートからの学生行動モデルの構築と考察”, 情報処理学会論文誌, 情報処理学会論文誌 51 巻, 4 号, 1215-1226 (2010)

[5] 佐藤和彦, 大川輝人, “事前対応型の修学指導支援システムの提案”, 電子情報通信学会技術研究報告. ET, 教育工学 107 巻, (205 号, pp57-60 (2007)

[6] 伊藤圭介, “データマイニングによる要注意学生の発見に関する研究”, 平成 25 年度名古屋工業大学修士論文 (2013)

[7] 石井一夫, “図解よくわかるデータマイニング”, 日刊工業新聞社 (2004)

[8] 鈴木護, “ベイジアンネットワーク入門”, 培風館 (2009)

[9] 田中和之, “ベイジアンネットワークの統計的推論の数理”, コロナ社 (2009)

[10] 木村陽一,“ベイジアンネットワーク：入門からヒューマンモデリングへの応用まで”, 行動計量学会セミナー資料 (2004)